http://www.hsdn.org
E-mail: info@hsdn.org

Принцип работы booru граббера

1. Скачивание XML листов
При скачивании XML листов производится подсчет их количества (каждый лист обычно имеет длину 1000 постов), после чего начинается скачивание листов (страниц). При необходимости можно задать диапазон страниц (например, с 1 по 3), таким образом скачивать не все листы, а только последние. Листы скачиваются в виде XML файлов, по одному на страницу. Файлы помещаются в специальную директорию.

2. Фильтрация XML листов
При проведении этой процедуры производится исключение уже имеющихся постов из листов, а также постов с запрещенными параметрами (тегами, рейтингом, размером и т.д.).
Фильтрация производится в 3 этапа:
1) Читается директория XML листов постов и строится их список
2) Читается каждый XML файл из списка
3) Производится проверка хешей постов на существование таковых в базе данных постов
4) Производится проверка параметров постов на допустимость
5) Данные о не отсеянных постах заносятся в читаемый XML файл, перезаписывая его
По окончанию процесса получаются отфильтрованные XML листы, которые затем можно передать в операцию скачивания постов.

3. Скачивание постов
Для проведения этой операции читается директория XML листов (которая содержит уже отфильтрованные листы), после чего по очереди читается каждый XML файл. После прочтения XML файла, производится скачивание указанных в нем постов. При скачивании проверяется также наличие скачиваемого файла во временной директории, это позволяет исключить повторное скачивание уже скачанных файлов.

4. Загрузка постов
Загрузка скаченных файлов осуществляется непосредственно через user-end интерфейс сайта. При проведении этой операции, как и в предыдущем случае, читается директория XML листов, после чего по очереди читается каждый XML файл, причем с последнего до первого. Такой порядок чтения важен, чтобы сохранить упорядоченность постов. После прочтения каждого XML файла, производится загрузка скаченных постов, проверяя наличие файла во временной директории, а также статус работы сайта.

5. Скачивание тегов
После успешной загрузки постов требуется перемаркировать (добавить) теги, связанные с этими постами. Теги хранятся в отдельных XML листах, которые требуется скачать. Скачивание производится аналогично, как и случае с XML листами тегов, за исключением того, фильтрации не происходит.

6. Загрузка тегов
После того, как листы тегов были скачены, происходит их чтение и занесение данных в базу данных тегов сайта. Для проведения этой операции читается директория XML листов тегов, после чего по очередности читается каждый файл, обновляя информацию в базе данных тегов.

Определения

Теги – специальные ключевые слова, используемые для связи постов
Пост – картинка (видео, и другая информация), сопровожденная тегами и другими служебными данными.
XML лист – специально структурированный XML файл (XML страница), содержащий информацию о присутствующих на сайте постов или тегов.
LRG – специализированный продукт, представляющий собой структурированную систему выборки и скачивания информации с удаленных сайтов.

Хранение конфигураций и структура задач

В случае, если применяется структурная модель LRG, задачи представляют собой отдельные классы (без описания рабочего функционала), расширенные родительскими классами, в которых описан весь функционал работы. Будет выделено два вида задач: работа с тегами и работа с постами. На каждый из этих видов будет применен свой родительский класс с описанием соответствующего функционала.

Структурная модель устройства задач (LRG-based)