

УЧЕБНЫЙ ЦЕНТР «ТОПЭКСПЕРТ»



ТОПЭКСПЕРТ

УЧЕБНЫЙ ЦЕНТР

Модуль проверки кода всех важных страниц

10 поток профессионального курса по SEO

Работу выполнил: Алексей Кротов

Дипломный руководитель: Дмитрий Иванов

Модуль: Проверка кода ответа всех важных страниц.

#1 Описание проекта

Используя:

- список всех важных внутренних страниц сайта важность документов определяется:
 - 1) дополнительным параметром “important=1” в СуБД для таблицы URL,
 - 2) из ручного списка внутренних URL сайта в .txt файле,
 - 3) конкретно для заданного внутреннего URL
 - 4) анализ приоритета страниц в Sitemap для добавления важности страниц

Необходимо:

- определить валидность URL
- определить код ответа данных страниц

Результат:

- вывод на экран админ-панели модуля список важных страниц с характерным для него кодом ответа сервера

примеры:

<http://vk.com/feed> код 200

[http:// vk.com /gfufg](http://vk.com/gfufg) код 404

<http://www.vk.com/> код 302

#2 Задачи, которые решает модуль

- информирует веб-мастера о правильности кодов ответа серверов важных (продвигаемых) страниц

#3 Входные данные (описание и таблица)

Обработка URL происходит на основе загрузки .txt файла, содержащего URL важных страниц (по одному URL на строчку), либо на основании СуБД с дополнительным параметром для каждой страницы- «important=1», который добавляется к странице при создании новых страниц, или же изменяется вручную. Данный параметр приписывается к таблице «URL» или же другим таблицам, содержащие сведения об URL (например: «Страницы») и не зависит от других строк в таблице.

Автор: Алексей Кротов, 10-ый поток курсов ТопЭксперт

Дипломный руководитель: Дмитрий Иванов

Таблицы примеров:

1)

Название раздела	Название категории	Название карточки	URL раздела	URL категории	URL карточки	important
Компьютерная техника			Hi-tech			1
Компьютерная техника	Нетбуки		Hi-tech	netbook		0
Компьютерная техника	Нетбуки	Asus UX31a Zenbook	Hi-tech	netbook	asus-ux-31-a	1

2)

Название страницы	URL страницы	important
Главная страница	/	1
Каталог	/catalog/	1
Самсунг Ж-666	/catalog/tv/Samsung-g666	0

3)

URL страницы	important
/catalog/	1
/catalog/tv/Samsung-g666	0

#4 Выходные данные

На выходе - на экране админ-панели модуля показывается список страниц, с характерными для них на момент сканирования кодом ответа сервера.

#5 Возможные кодов ответа сервера

1xx: Informational (информационные):

100 Continue («продолжить»)^{[1][3]}.

101 Switching Protocols («переключение протоколов»)^{[1][3]}.

102 Processing («идёт обработка»).

105 Name Not Resolved («не удастся преобразовать DNS-адрес сервера»).

Автор: Алексей Кротов, 10-ый поток курсов ТопЭксперт

Дипломный руководитель: Дмитрий Иванов

2xx: Success (успешно):

- 200 OK («хорошо»)^{[1][3]}.
- 201 Created («создано»)^{[1][3][4]}.
- 202 Accepted («принято»)^{[1][3]}.
- 203 Non-Authoritative Information («информация не авторитетна»)^{[1][3]}.
- 204 No Content («нет содержимого»)^{[1][3]}.
- 205 Reset Content («сбросить содержимое»)^{[1][3]}.
- 206 Partial Content («частичное содержимое»)^{[1][3]}.
- 207 Multi-Status («многостатусный»)^[5].

226 IM Used («использовано IM»).

3xx: Redirection (перенаправление):

- 300 Multiple Choices («множество выборов»)^{[1][6]}.
- 301 Moved Permanently («перемещено навсегда»)^{[1][6]}.
- 302 Moved Temporarily («перемещено временно»)^{[1][6]}.
- 302 Found («найдено»)^[6].
- 303 See Other (смотреть другое)^{[1][6]}.
- 304 Not Modified (не изменялось)^{[1][6]}.
- 305 Use Proxy («использовать прокси»)^{[1][6]}.
- 306 — *зарезервировано* (код использовался только в ранних спецификациях)^[6].
- 307 Temporary Redirect («временное перенаправление»)^[6].

4xx: Client Error (ошибка клиента):

- 400 Bad Request («плохой, неверный запрос»)^{[1][3][4]}.
- 401 Unauthorized («неавторизован»)^{[1][3]}.
- 402 Payment Required («необходима оплата»)^{[1][3]}.
- 403 Forbidden («запрещено»)^{[1][3]}.
- 404 Not Found («не найдено»)^{[1][3]}.
- 405 Method Not Allowed («метод не поддерживается»)^{[1][3]}.
- 406 Not Acceptable («неприемлемо»)^{[1][3]}.
- 407 Proxy Authentication Required («необходима аутентификация прокси»)^{[1][3]}.
- 408 Request Timeout («истекло время ожидания»)^{[1][3]}.
- 409 Conflict («конфликт»)^{[1][3][4]}.
- 410 Gone («удалён»)^{[1][3]}.
- 411 Length Required («необходима длина»)^{[1][3]}.
- 412 Precondition Failed («условие ложно»)^{[1][3][7]}.
- 413 Request Entity Too Large («размер запроса слишком велик»)^{[1][3]}.
- 414 Request-URI Too Large («запрашиваемый URI слишком длинный»)^{[1][3]}.
- 415 Unsupported Media Type («неподдерживаемый тип данных»)^{[1][3]}.
- 416 Requested Range Not Satisfiable («запрашиваемый диапазон не

достижим»)^[3].

417 Expectation Failed («ожидаемое неприемлемо»)^[3].

422 Unprocessable Entity («необрабатываемый экземпляр»).

423 Locked («заблокировано»).

424 Failed Dependency («невыполненная зависимость»).

425 Unordered Collection («неупорядоченный набор»)^[8].

426 Upgrade Required («необходимо обновление»).

428 Precondition Required («необходимо предусловие»)^[9].

429 Too Many Requests («слишком много запросов»)^[9].

431 Request Header Fields Too Large («поля заголовка запроса слишком большие»)^[9].

449 Retry With («повторить с»)^[2].

451 Unavailable For Legal Reasons («недоступно по юридическим причинам»)^[10].

456 Unrecoverable Error («некорректируемая ошибка»).

499 Используется Nginx, когда клиент закрывает соединение до получения ответа.

5xx: Server Error (ошибка сервера):

500 Internal Server Error («внутренняя ошибка сервера»)^{[1][3]}.

501 Not Implemented («не реализовано»)^{[1][3]}.

502 Bad Gateway («плохой, ошибочный шлюз»)^{[1][3]}.

503 Service Unavailable («сервис недоступен»)^{[1][3]}.

504 Gateway Timeout («шлюз не отвечает»)^{[1][3]}.

505 HTTP Version Not Supported («версия HTTP не поддерживается»)^{[1][3]}.

506 Variant Also Negotiates («вариант тоже проводит согласование»)^[11]

507 Insufficient Storage («переполнение хранилища»).

508 Loop Detected («обнаружена петля»)^[12]

509 Bandwidth Limit Exceeded («исчерпана пропускная ширина канала»).

510 Not Extended («не расширено»).

511 Network Authentication Required («требуется сетевая аутентификация»)^[9].

#6 Формулы

Номер	Условие
1	Проверка валидности URL адресов (для .рф зоны и русскоязычных доменов – использовать пуникод-кодировку, проверка протокола соединения (проверяем протокол = HTTP\HTTPS)
2	При работе 3xx кода ответа для URL ответ вида: URL1 код 3xx → URL2 код 200
3	Все полученные данные на выходе не содержат заглавных букв

#7 Карта логических связей выполнения модуля



#8 Список модулей, с которыми взаимодействует модуль

- модуль создания страниц
- группа модулей перелинковки
- модуль генерации карты сайта
- модуль поиска битых ссылок и ссылок на редиректы
- модуль формирования бекапов
- модуль проверки корректности перелинковки

#9 Описание процессов взаимодействия

1. Модуль создания страниц
 - при создании новых страниц можно явно указывать, что данная страница является важной, при этом в СУБД будет записываться параметр *important=1*

Группа модулей перелинковки:

Перелинковка для повышение анкорного ранжирования

Перелинковка для индексации новых страниц

Перелинковка для индексации старых страниц

Перелинковка для повышения статического веса

Перелинковка для увеличения внутреннего трафика)

- при наличии длительного ответа страниц с кодом 4xx исключать страницы из перелинковки, исключая возможность ссылаться другие страницы на нерабочую
2. Модуль генерации карты сайта
 - при наличии длительного ответа страниц с кодом 5xx исключить их из Sitemap

Автор: Алексей Кротов, 10-ый поток курсов ТопЭксперт

Дипломный руководитель: Дмитрий Иванов

3. Модуль поиска битых ссылок и ссылок на редиректы
 - при наличии 4xx и 3xx внеплановая активация модуля
4. Модуль формирования бекапов
 - при длительном ответе 5xx –предлагает восстановить последний работоспособный бэкап
5. Модуль проверки корректности перелинковки
 - при наличии 4xx кодов ответа внеплановая активация модуля

#10 Предполагаемая нагрузка

- при проверке важных страниц нагрузка будет средняя, линейно зависящая от количества важных страниц. Путем применения периодичности запросов к серверу, мы можем минимизировать нагрузку на сервер, при этом увеличится общее время выполнения задачи.

#11 Особые требования

6. Ручная проверка конкретных страниц
7. Ручная пакетная проверка конкретных страниц
8. Автоматизированная по запросу проверка страниц из СУБД
9. Возможность изменять важность страниц (для работы модуля из СУБД)
10. Возможность проверки всех, без исключения страниц сайта

#12 Процесс остановки модуля

- в случае, если режим проверки заголовков модуля превысил допустимое время (рекомендуется установить время ожидания разового соединения с сервером >~20 сек, при большом кол-ве важных страниц (> 10 000 страниц))
- в случае сбора и вывода информации по всем, интересующим нас важным страницам.

#13 Процесс запуска модуля

- модуль запускается в ручную, при этом пользователь может выбрать 3 варианта запуска модуля: для конкретной страницы, пакетная загрузка .txt файла с URL важных страниц, работа с СУБД с *“important=1”*
- автоматически с помощью Cron в настройках Crontab. Интервал времени с 8:00-до 23:00 – 15 минут, с 23:00 до 8:00 – 3 минуты. В случае автоматического запуска используется работа СУБД с параметром страниц *“important=1”*.

#14 Предполагаемое расширение модуля

Расширение не планируется

#15 Возможные причины поломки модуля

- выход из строя БД, содержащей параметра "*important*", отсутствие хотя бы одного значения "*important=1*",

#16 Работа модуля в случае поломки

- При анализе СУБД таблицы URL страниц и отсутствии параметра *important* или не обнаружении параметра *important=1* хотя бы для одной страницы выводится сообщение в админ-панели модуля и отправкой e-mail веб-мастеру (группе веб-мастеров) о том, что важные страницы не назначены.
- При работе модуля через конкретную URL-страницу или пакетную загрузку через .txt файл возможность поломки модуля --> 0.