ТОПЭКСПЕРТ

УЧЕБНЫЙ ЦЕНТР

Модуль перелинковки для индексации страниц

10 поток профессионального курса по SEO

Работу выполнил: Рамиль Алиякберов

Дипломный руководитель: Дмитрий Иванов

ТОПЭКСПЕРТ
УЧЕБНЫЙ ЦЕНТР

Техническое задание на модуль перелинковки для индексации страниц

Оглавление

Автор: Рамиль Алиякберов, 10-ый поток курсов ТопЭксперт

Дипломный руководитель: Дмитрий Иванов

Описание

Модуль перелинковки для индексации страниц предназначен для того, чтобы обеспечить быструю индексацию новых и повышения вероятности индексации существующих но не проиндексированных страниц.

Порядок работы

Модуль должен работать в 3-х режимах:

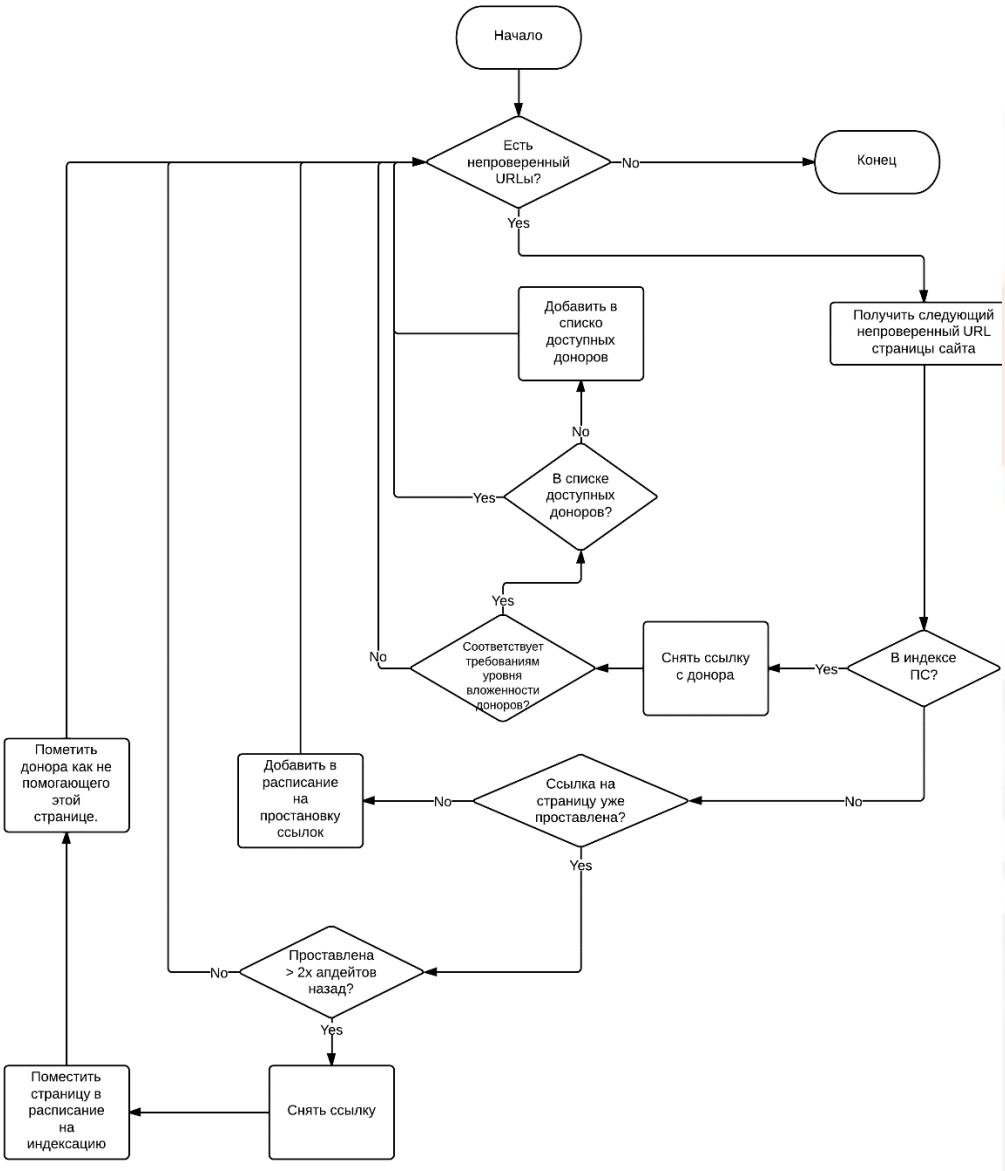
1. Автоматическое
    a. Отслеживание страниц и добавление/снятие ссылок на них
    b. Привязка анкоров к страницам акцепторам
    c. Создание индекса акцептор - доноры
    d. Постановка в очередь на простановку новой страницы при ее добавлении.
2. Ручной запуск

Подсистема анализа

Подсистема запускается в соответствии с настройками (ежечасно/ежедневно/еженедельно и т.п.) либо в ручную и проверяет наличие страниц в индексе поисковой системы (ПС). В случае, если страница отсутствует в индексе ПС, а так же ее нет в очереди на простановку и ссылка на нее еще не размещена, модуль добавляет страницу в очередь на простановку ссылки на эту страницу. В случае, если страница есть в индексе ПС и она удовлетворяет условиям донорства, то данная страница добавляется в очередь доноров для простановки.

Автор: Рамиль Алиякберов, 10-ый поток курсов ТопЭксперт

Дипломный руководитель: Дмитрий Иванов



Автор: Рамиль Алиякберов, 10-ый поток курсов ТопЭксперт

Дипломный руководитель: Дмитрий Иванов

Подсистема привязки анкоров к акцепторам

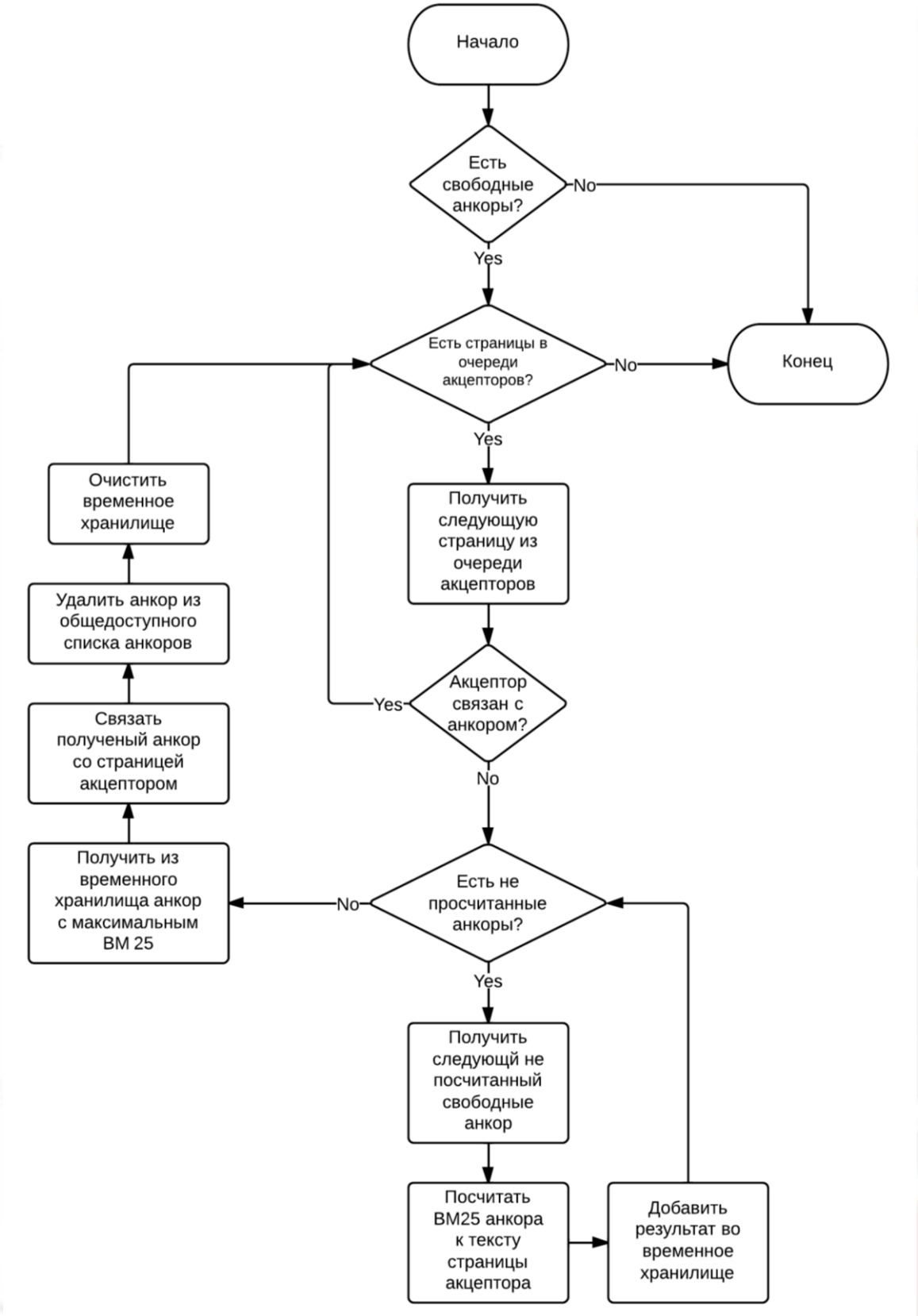
Подсистема создается для того, чтобы в автоматическом режиме выбирать наиболее подходящий акцептору анкор, который будет проставлен на страницу донора.

Порядок работы:

1. Подсистема проверяет наличие в очереди акцепторов которым еще не был присвоен анкор.
2. Проверяет наличие анкоров доступных для привязки
3. Привязывает а акцептору анкор с максимальным значением BM25

Автор: Рамиль Алиякберов, 10-ый поток курсов ТопЭксперт

Дипломный руководитель: Дмитрий Иванов



Автор: Рамиль Алиякберов, 10-ый поток курсов ТопЭксперт

Дипломный руководитель: Дмитрий Иванов

Подсистема индексации акцептор – доноры

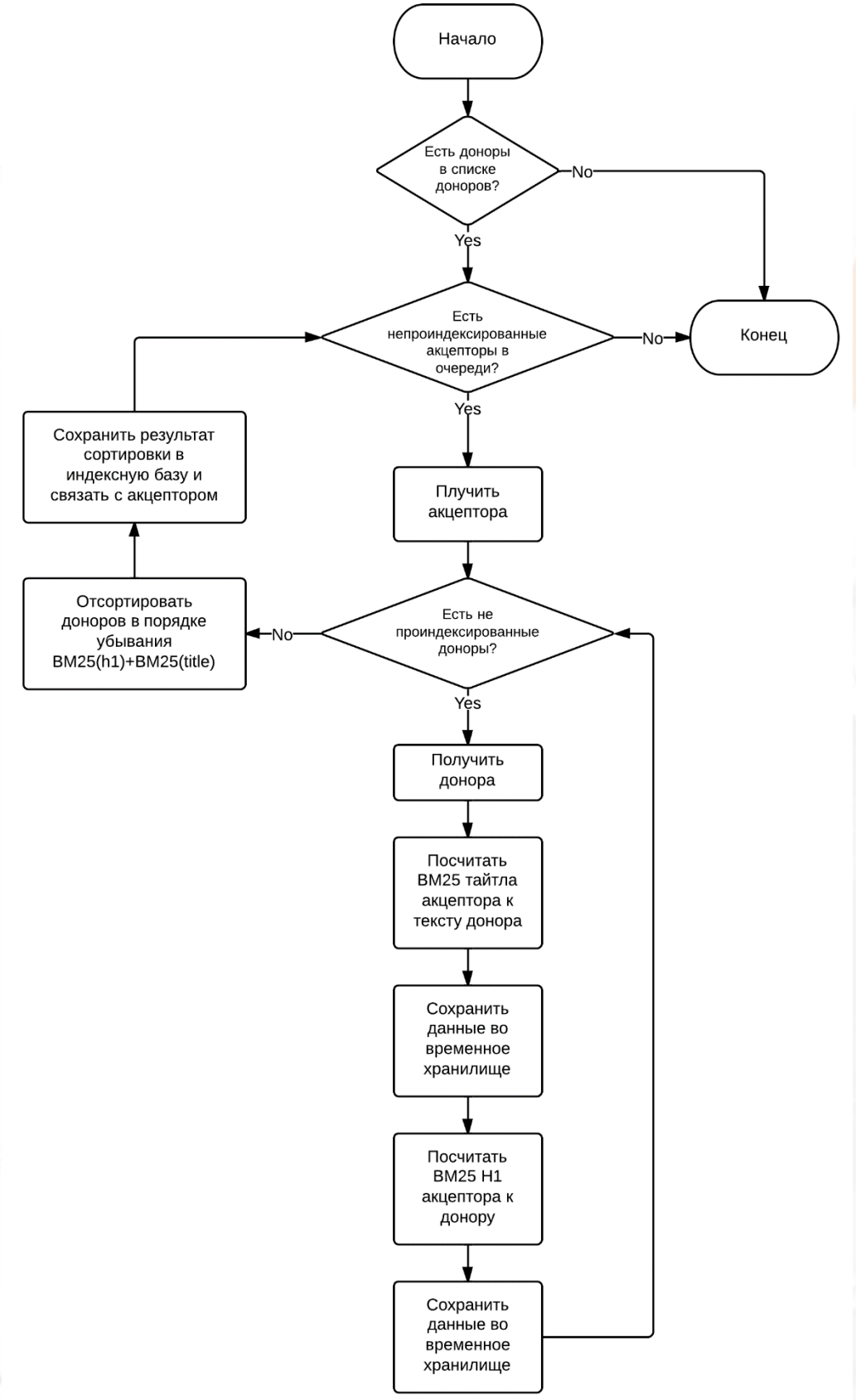
Назначение подсистемы – построить индексную базу акцептор-доноры, где в верху списка доноров размещен наиболее подходящий данному акцептору и далее в порядке убывания степени соответствия.

Порядок работы:

1. Подсистема проверяет наличие акцепторов для которых не был составлен индекс
2. Проверяет наличие доступных доноров
3. Просчитывает BM25 для h1 и title акцептора по отношению к тексту страницы донора
4. Упорядочивает доноров в порядке уменьшения ранга BM25
5. Сохраняет полученный результат в индексную базу

Автор: Рамиль Алиякберов, 10-ый поток курсов ТопЭксперт

Дипломный руководитель: Дмитрий Иванов



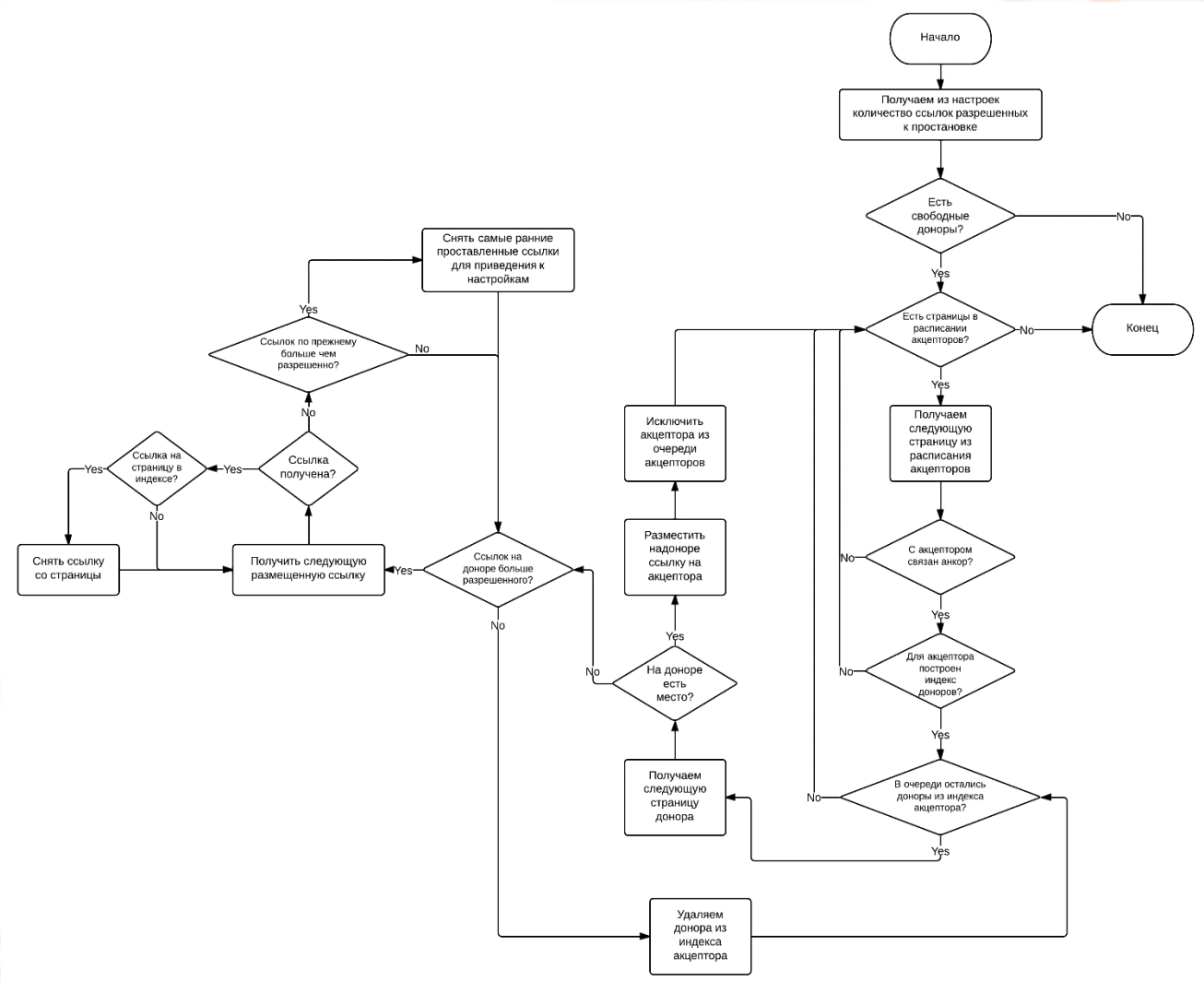
Автор: Рамиль Алиякберов, 10-ый поток курсов ТопЭксперт

Дипломный руководитель: Дмитрий Иванов

Подсистема простановки

Данная подсистема запускается в соответствии с настройками (ежечасно/ежедневно/еженедельно и т.п.) либо в ручную. Работа подсистемы состоит из следующих шагов:

1. Получить список страниц доноров, доступных к простановке
2. Получить из настроек количество ссылок, разрешенных к простановке на одной странице донора.
3. Проверяет, есть ли в расписании акцепторы требующие простановки
4. Проверяет соответствие акцептора двум условиям:
    a. Акцептор связан с анкором
    b. Для акцептора построен индекс акцептор-доноры
5. Получает следующую доступную страницу донора для акцептора, в соответствии с построенным индексом
6. Проверяет наличие места на доноре. В случае его отсутствия удаляет донора из индекса акцептора и переходит к шагу 5
7. Если место на доноре есть, то подсистема размещает на нем ссылку на акцептора с привязанным к акцептору анкором
8. Удаляет акцептора из очереди



Автор: Рамиль Алиякберов, 10-ый поток курсов ТопЭксперт

Дипломный руководитель: Дмитрий Иванов

Настройки модуля

Основные настройки

1. Указание глубины вложенности страниц допускающихся к добавлению в списко доноров
2. Ввод списка анкоров для ссылок
3. Выбор ПС для проверки индексации

Подсистема анализа

1. Указание интервала запуска модуля
    a. Ежечасно
    b. Ежедневно
    c. Еженедельно
    d. Ежемесячно

Подсистема привязки анкоров

1. Указание интервала запуска модуля
    a. Ежечасно
    b. Ежедневно
    c. Еженедельно
    d. Ежемесячно

Подсистема создания индекса акцептор-доноры

1. Указание интервала запуска модуля
    a. Ежечасно
    b. Ежедневно
    c. Еженедельно
    d. Ежемесячно

Подсистема простановки

1. Указание интервала запуска модуля
    a. Ежечасно
    b. Ежедневно
    c. Еженедельно
    d. Ежемесячно
2. Указание количества ссылок разрешенных для простановки на доноре

Управление модулем

Основные настройки

1. Показать форму для ввода настроек:
    a. Макс глубина вложенности для доноров – поле ввода, число
    b. Ввод списка анкоров – текстовое поле. Правило ввода – каждый новый анкор на новой строке
    c. Выпор ПС для проверки – выпадающий список. Варианты:
        i. Yandex
        ii. Google

Управление очередью простановки

Вывод данных модуля:

Приоритет	URL страницы	Дата создания	Количество	Привязанный	Доноры

Автор: Рамиль Алиякберов, 10-ый поток курсов ТопЭксперт

Дипломный руководитель: Дмитрий Иванов

размещения			символов на странице	анкор	
1 ⇕	Site.ru/page1/	01/01/2000	1000	A1 ✚ ✖	Изменить
2 ⇕	Site.ru/page2/	01/02/2000	1025	A2 ✚ ✖	Изменить
3 ⇕	Site.ru/page3/	02/01/2000	998	-- ✚	Изменить
…	…	….	…	…	…
N ⇕	Site.ru/pageN/	01/01/2013	20	-- ✚	Изменить

С помощью стрелок рядом с цифрой приоритета, можно менять приоритет, так же, должна быть возможность задать номер приоритета в ручную.

«--» - означает, что анкор к акцептору еще не привязан.

С помощью ✖ можно отвязать автоматически привязанный анкор.

С помощью ✚ можно добавить анкор для страницы вручную.

Чем ниже цифра приоритета, тем раньше ссылка на данную страницу будет размещена

Управление индексом доноров акцептора

Данная страница параметров открывается для конкретного акцептора и позволяет изменять приоритет (ранг) донора в индексной базе акцептора

Приоритет размещения (ранг соответствия акцептору)	URL страницы	Размещенных ссылок	Разрешено к размещению	Удалить из индекса
1 ⇕	Site.ru/page1/	1	5	Удалить
2 ⇕	Site.ru/page2/	10	10	Удалить
3 ⇕	Site.ru/page3/	3	0	Удалить
…	…	….	…	…
N ⇕	Site.ru/pageN/	6	0	Удалить

⇕ - служит для изменения ранга донора относительно акцептора.

Так же должна быть возможность вручную задать ранг донора

Кнопка Удалить служит для удаления донора из индекса акцептора.

Управление списком доноров

Приоритет размещения	URL страницы	Размещенных ссылок	Разрешено к размещению
1	Site.ru/page1/	1	5
2	Site.ru/page2/	10	10
3	Site.ru/page3/	3	0
…	…	….	…
N	Site.ru/pageN/	6	0

Автор: Рамиль Алиякберов, 10-ый поток курсов ТопЭксперт

Дипломный руководитель: Дмитрий Иванов

Таблица позволяет управлять списком доноров.

Так же, помимо глобальных настроек количества разрешенных к размещению ссылок, для каждой страницы можно задать свое количество.

Автор: Рамиль Алиякберов, 10-ый поток курсов ТопЭксперт

Дипломный руководитель: Дмитрий Иванов

Формулы

BM25

Пусть дан запрос Q, содержащий слова $q_1, \cdots, q_n$, тогда функция BM25 даёт следующую оценку релевантности документа D запросу Q:

$$\text{score}(D,Q) = \sum_{i=1}^{n} \text{IDF}(q_i) \cdot \frac{f(q_i, D) \cdot (k_1 + 1)}{f(q_i, D) + k_1 \cdot (1 - b + b \cdot \frac{|D|}{\text{avgdl}})},$$

где $f(q_i, D)$ есть частота слова (*term frequency, TF*) q_i в документе D, $|D|$ есть длина документа (количество слов в нём), а $avgdl$ — средняя длина документа в коллекции. k_1 и b — свободные коэффициенты, обычно их выбирают как $k_1 = 2.0$ и $b = 0.75$.

$\text{IDF}(q_i)$ есть обратная документная частота (*inverse document frequency, IDF*) слова q_i. Есть несколько толкований IDF и небольших вариации его формулы. Классически, она определяется как:

$$\log \frac{N}{n(q_i)},$$

Пример реализации BM25 с помощью Excel есть во вложении (LinksBM25.xlsm)

Определение соответствия доноров и анкора акцептору

Соответствие анкора вычисляется по простой формуле BM25 score(D,Q), где D – текст страницы акцептора, а Q – текст анкора.

Соответствие донора акцептору вычисляется как score(D, title)+score(D, h1), где D – текст страницы донора, title – meta-title страницы акцептора, и h1 – h1 страницы акцептора.

Автор: Рамиль Алиякберов, 10-ый поток курсов ТопЭксперт

Дипломный руководитель: Дмитрий Иванов

Модули, с которыми может взаимодействовать данный модуль

1. Генератор ЧПУ, для получения ЧПУ ссылок для размещения
2. Модуль простановки rel=canonical, для выявления канонической страницы
3. Модули генерации и проверки Title и H1
4. Другие модули перелинковки, для вычисления максимальных объемов простановки для каждого донора
5. Генератор анкоров, для пополнения списка анкоров
6. Другие модули повышения индексации, для получения/отправки сведений о проиндексированных и не проиндексированных страницах
7. Модуль анализа путей роботов по логам, для выявления периодичности захода роботов на определенные разделы и подстройки графика запуска конкретных подсистем в соответствии с этими данными
8. Модуль формирования бэкапов

Предполагаемая нагрузка

Предполагается высокая нагрузка при наличии на сайте большого количества страниц.

Настраивать запуск модуля желательно на то время, когда посещаемость сайта минимальна. Желательно не запускать отдельные подсистемы совместно. Только по очереди, в следующем порядке запуска:

1. Проверка индексации (3)
2. Привязка анкоров (2)
3. Построение индекса акцептор-доноры (4)
4. Простановка (1)

*В скобках проставлен предполагаемый ранг нагрузки на сервер. Цифры теоритические. Реальные нагрузки будут зависеть от объемов списков доноров и акцепторов.

Бэкапы

Перед запускам каждой из подсистем, модуль должен делать резервную копию уже обработанных данных. На диске, а лучше на отдельном бэкап сервере, должна сохраняться копия таблиц и данных Модуля.

Восстановление происходит путем развертывания сохраненных копий в БД сайта.

Логирование

При запуске каждой подсистемы автоматически создается лог файл с сообщениями о ходе работы подсистемы. При каждом запуске создается новый файл. Хранится максимум 5 лог-файлов для каждой подсистемы.

В лог файл должны писаться все предупреждения и ошибки в работе подсистем, как фатальные, так и не фатальные.

В случае выявления не фатальной ошибки в работе одной из подсистем эта информация, помимо записи в лог, должна добавляться в уведомления администратору сайта.

Работа в случае поломки

В случае поломки на этапе работы любой из подсистем модуля, работа модуля останавливается. Все подсистемы переводятся в статус PENDING, вызвавшая сбой подсистема переводится в статус ERROR. В

статусах ERROR и PENDING работа модуля приостанавливается независимо от расписания запуска. Работа возобновляется только после перевода всех подсистем модуля в статус OK.

Так же, в случае поломки одной из подсистем, должна быть автоматически развернута последняя резервная копия данных вызвавшей ошибку подсистемы.

Автор: Рамиль Алиякберов, 10-ый поток курсов ТопЭксперт

Дипломный руководитель: Дмитрий Иванов