

Санкт-Петербургский государственный университет  
информационных технологий, механики и оптики

Факультет информационных технологий и программирования  
Кафедра компьютерных технологий

Лебедева А. В.

## Статическая приоритезация поискового робота.

Дипломная работа

Научный руководитель: Романенко А. А.

Санкт-Петербург  
2013

# Содержание

Содержание . . . . .	3
Введение . . . . .	5
<b>Глава 1. Обзор предметной области . . . . .</b>	<b>7</b>
1.1 Основные понятия . . . . .	7
1.1.1 Интернет как граф . . . . .	7
1.1.2 Поисковая система . . . . .	7
1.2 Оценка важности веб-страниц . . . . .	8
1.2.1 Алгоритмы ранжирования, использующие свойства веб-графа . . . . .	9
1.2.2 Алгоритмы, использующие текстовое представление веб-страниц . . . . .	11
1.3 Поисковый робот . . . . .	11
1.3.1 Общее описание . . . . .	11
1.3.2 Обязательные свойства . . . . .	11
1.3.3 Желательные свойства . . . . .	12
1.4 Алгоритм приоритезации поискового робота . . . . .	14
1.4.1 Стратегии отбора . . . . .	14
1.4.2 Стратегии повторного обхода . . . . .	16
<b>Глава 2. Предлагаемый метод . . . . .</b>	<b>17</b>
2.1 Стратегия приоритезации . . . . .	17
2.2 Этапы алгоритма . . . . .	18
2.3 Статические особенности URL-адреса . . . . .	20
<b>Глава 3. Экспериментальные результаты . . . . .</b>	<b>22</b>
3.1 Реализация . . . . .	22
3.2 Параметры запусков . . . . .	25

3.3 Результаты . . . . .	26
<b>Заключение . . . . .</b>	<b>27</b>
<b>Источники . . . . .</b>	<b>29</b>
<b>Приложения . . . . .</b>	<b>30</b>
А. Программный код . . . . .	30

# Введение

Объем информации в интернете в последнее время очень быстро растет. На данный момент число сайтов исчисляется десятками миллиардов, а число веб-страниц — десятками миллиардов. Кроме того, информация в интернете быстро изменяется, веб-страницы обновляются все чаще.

В связи с данными особенностями развития интернета поиск необходимой информации является очень сложной задачей для пользователя. Для упрощения этой задачи были созданы поисковые системы.

Поисковый робот является частью поисковой системы, необходимой для ее корректного функционирования. Его задачей является перебор страниц интернета. Загруженные роботом страницы сохраняются в базу данных поисковой системы для последующих выдач в пользовательских поисковых запросах.

Чтобы удовлетворять требованиям пользователей, в поисковая система должна хранить в своей базе веб-страницы, которые будут востребованы пользователями поисковой системы. Для этого необходима стратегия отбора веб-страниц, которые поисковый робот должен скачать, и в каком порядке. Для чего необходим алгоритм ранжирования веб-страниц, который будет определять важность веб-страницы. Существует два основных подхода к ранжированию веб-страниц. Первый подход основан на анализе веб-графа. Вторым — на анализе текстовых особенностей веб-страницы, особенностей ее URL-адреса и анализе ее содержимого.

На основе анализа веб-графа построен алгоритм приоритизации PageRank, который успешно использует поисковая система Google. Но вычисление подобного ранга для каждой страницы является очень ресурсоем-

кой операцией, поскольку для этого необходимо использовать информацию о всем веб-графе, размер которого велик.

На базе графовых алгоритмах ранжирования построен ряд стратегий отбора веб-страниц. Например, OPIC [1], который также является аппроксимацией вычисления PageRank. Также были разработаны алгоритмы, использующие в основе приоритезации информацию о расстоянии между страницами в веб-графе, например, FICA [2].

На основе анализа статических факторов страницы был разработан алгоритм ранжирования веб-страниц FRank, описанный в работе [3]. Но вопрос его применения к приоритезации поискового робота остается открытым, поскольку оценка важности страницы требует информации не только о URL-адресе веб-страницы, но и об ее содержимом.

Кроме того, поисковая система должна располагать актуальной информацией о состоянии веб-страниц. Для чего поисковый робот должен повторно скачивать веб-страницы, уже имеющиеся в базе поисковой системы. Здесь возникает вопрос: как часто повторно скачивать веб-страницы, и в каком порядке. Для чего необходима стратегия повторного скачивания.

Алгоритм приоритезации поискового робота представляет собой комбинацию двух стратегий: отбора веб-страниц и повторного скачивания.

В данной работе описывается разработка программы, реализующей поискового робота. Для алгоритма приоритезации которого применяется как анализ веб-графа, так и анализ текстовых особенностей URL-адресов веб-страниц.

# Глава 1. Обзор предметной области

В данной главе приведено описание предметной области. Описаны подходы к алгоритмам приоритезации поискового робота.

## 1.1. ОСНОВНЫЕ ПОНЯТИЯ

### 1.1.1. Интернет как граф

Одной из формальных моделей, с помощью которой можно представить интернет, является его представление в виде направленного графа. Вершинами в этом графе являются веб-страницы, а ребрами — ссылки. Из вершины  $u$  есть ребро в вершину  $v$ , если на веб-странице, соответствующей узлу  $u$  есть ссылка на веб-страницу, соответствующей вершине  $v$ . Заметим, что данный граф является динамическим, поскольку веб-страницы часто изменяются, какие-то из них удаляются, а также создаются новые.

### 1.1.2. Поисковая система

Поисковая система — это программный комплекс, предназначенный для помощи пользователю осуществления поиска информации в интернете. Основные компоненты поисковой системы представлены на рисунке 1.1.

Когда пользователь осуществляет информационный запрос, поисковая система формирует упорядоченное подмножество веб-страниц, хранящихся в ее индексе. Таким образом, качество работы поисковой системы во многом зависит от содержимого индекса, формирование которого осуществляется индексатором из множества веб-страниц коллекции, полученной в процессе обхода интернета, потому очень важно, чтобы в коллекции содержались важные веб-страницы, то есть веб-страницы, востребованные пользователем. Алгоритмы определения важности веб-страниц описаны в разделе 1.2. Коллекцию, в свою очередь, формирует поисковый робот. По-

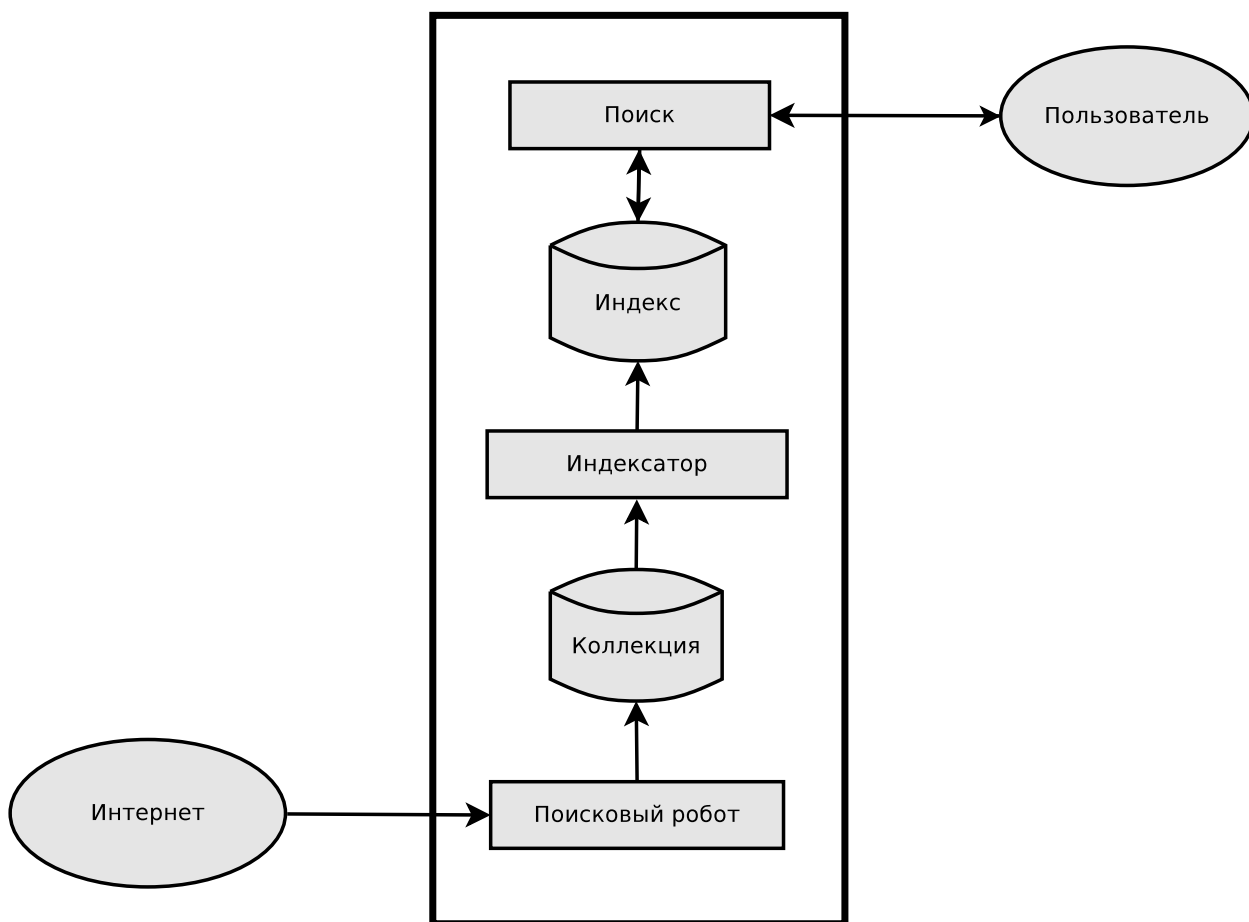


Рис. 1.1. Основные компоненты поисковой системы

дробное описание работы которого приведено в разделе 1.3. Таким образом, поисковый робот в процессе обхода должен производить оценку страниц, и скачивать лишь важные.

## 1.2. ОЦЕНКА ВАЖНОСТИ ВЕБ-СТРАНИЦ

Оценка важности веб-страницы называется рангом. Процесс получения оценки важности страницы называется ранжированием. Рассмотрим существующие методы, осуществляющие ранжирование.

### 1.2.1. Алгоритмы ранжирования, использующие свойства веб-графа

Данная категория алгоритмов использует представление интернета в виде графа, описанное в разделе 1.1.1.

#### PageRank

Метод разработан Сергеем Брином и Ларри Пейджем, основателями поисковой системы Google. Первая статья, описывающая этот алгоритм ранжирования была опубликована в 1998 году [4]. Основная идея метода заключается в том, что страницы, чаще посещаемые в ходе случайного обхода веб-графа, имеют большую важность. Пусть имеется  $N$  веб-страниц. Обход начинается с веб-страницы  $p$ , и в каждый момент времени с вероятностью  $\alpha$  осуществляется переход по случайной ссылке, имеющейся на текущей странице. Также с вероятностью  $1 - \alpha$  мы переходим на случайную страницу. Итоговая формула для вычисления PageRank имеет следующий вид:

$$PageRank(p) = \frac{1 - \alpha}{N} + \alpha * \sum_{p' \in In(p)} \frac{PageRank(p')}{|Out(p')|}$$

где  $In(p)$  — множество страниц, содержащих ссылку на  $p$ ,  $Out(p')$  — множество страниц, на которые ссылается  $p'$ , а  $\alpha$  — фиксированный параметр ( $0 < \alpha < 1$ ).

Ранжирование с использованием этого алгоритма является достаточно ресурсоемкой операцией, и требует знания о всей структуре веб-графа, что делает применение данного подхода на практике достаточно затруднительным.

#### HITS

Hyperlink-Induced Topic Search (HITS) — метод, разработанный в 1999 году Джоном Клейнбергом и описанный в работе. В отличие от



*PageRank*, является запросо-зависимым алгоритмом ранжирования, то есть полученные в результате анализа ссылок оценки веб-страниц зависят от конкретного поискового запроса. Базовая идея алгоритма заключается в том, что определенные веб-страницы являются *посредниками* (*hubs*), то есть используются как каталоги, указывающие на другие страницы, являющиеся авторитетными источниками (*authorities*) или *авторами* [5], которые содержат необходимую пользователю информацию. Тогда страница, указывающая на большое количество хороших *авторов*, является хорошим *посредником*, и наоборот, страница, на которую ссылается много хороших *посредников*, является хорошим *автором*. Основываясь на этом, для каждой страницы вычисляется две оценки: оценка авторитетности и посредническая оценка.

Для начала работы алгоритма необходимо получить множество наиболее релевантных страниц для текущего поискового запроса, называемое *корневым множеством* (*root set*). На его основе формируется *базовое множество* (*base set*), получаемое в результате добавления к *корневому множеству* всех веб-страниц, связанных с ним ссылками, как исходящими, так и входящими. Все вычисления проводятся на подграфе, сформированном *базовым множеством*. Основной алгоритм выполняет ряд итераций, на каждой из которых происходит обновление оценок для каждой вершины подграфа. Обновление посреднической оценки выполняется посредством суммирования оценок авторитетности каждой из вершин, на которые указывает текущая. В свою очередь, обновление авторитетной оценки осуществляется путем суммирования посреднических оценок вершин, ссылающихся на текущую. Окончательные оценки формируются после выполнения большого числа итераций алгоритма.

### **1.2.2. Алгоритмы, использующие текстовое представление веб-страниц**

Методы данной категории используют текстовое представление страниц.

## **1.3. ПОИСКОВЫЙ РОБОТ**

### **1.3.1. Общее описание**

Одним из важнейших условий функционирования поисковой системы является осуществление обхода интернета (web crawling). Его производит поисковый робот (web crawler or spider). Принцип его работы заключается в следующем. Робот хранит очередь URL-адресов веб-страниц, которые ему необходимо обойти. На момент начала обхода в ней находится некоторое исходное множество адресов. На каждой итерации робот извлекает из очереди следующий URL-адрес и скачивает веб-страницу, которая ему соответствует. Затем полученная страница обрабатывается, из нее извлекается содержимое и ссылки. Содержимое передается индексатору для последующей обработки и добавления страницы в базу поисковой системы. А URL-адреса страниц, содержащихся в ссылках, извлеченных с текущей страницы, добавляются в очередь на скачивание.

В книге [6] сформулированы основные свойства поискового робота. Они разделены на две категории: обязательные и желательные.

### **1.3.2. Обязательные свойства**

Обязательными свойствами поискового робота являются:

1. Устойчивость.
2. Вежливость.

Остановимся на каждом из них подробнее.

## Устойчивость

Число страниц в интернете потенциально может быть бесконечно, поскольку некоторые веб-сервера генерируют динамические веб-страницы по запросу на их создание, которым, в частности, является переход по ссылке. Таким образом, поисковый робот может заиклиться на определенном хосте, бесконечно скачивая вновь сгенерированные динамические страницы. Такой механизм называется ловушкой для роботов (spider trap). Поисковый робот должен быть устойчив к подобного рода явлениям.

## Вежливость

Для сайтов существуют явные и неявные правила поведения поискового робота при работе с ними. Эти правила регулируют частоту обращения к хосту, а также возможность доступа к его содержимому. Многие сайты имеют в корне специальный файл *robots.txt*, содержащий желательные правила поведения. Поисковый робот должен соблюдать их.

### 1.3.3. Желательные свойства

Желательными свойствами поискового робота являются:

1. Распределенность.
2. Масштабируемость.
3. Производительность и эффективность.
4. Качество.
5. Свежесть.
6. Расширяемость.

Остановимся на каждом из них подробнее.

## **Распределенность**

Подразумевает возможность функционирования поискового робота одновременно на многих машинах.

## **Масштабируемость**

Подразумевает возможность увеличения эффективности поискового робота при увеличении ресурсов. Таких как увеличение числа машин или расширения полосы пропускания.

## **Производительность и эффективность**

Подразумевает эффективное использование поисковым роботом доступных ему ресурсов.

## **Качество**

Желательно, чтобы поисковый робот при скачивании веб-страниц отдавал предпочтение более качественным. Для выполнения этого свойства необходима стратегия приоритезации страниц. Существующие стратегии подробнее рассмотрены в разделе 1.4.

## **Свежесть**

Для хранения свежей информации о веб-страницах в базе поисковой системы необходимо, чтобы поисковый робот обходил ранее скачанные страницы с частотой, соответствующей частоте их обновления. Способы выполнения этого свойства также рассмотрены в разделе 1.4.

## **Расширяемость**

Поисковый робот должен быть разработан так, чтобы имела возможность расширения для решения новых задач. Таких как добавление новых протоколов передачи данных или новых форматов данных.

## 1.4. АЛГОРИТМ ПРИОРИТЕЗАЦИИ ПОИСКОВОГО РОБОТА

Под приоритезацией подразумевается введение полного или частичного порядка на множестве веб-страниц, которое поисковый робот должен обработать. Это необходимо для того, чтобы поисковая система хранила в базе страницы, содержащие информацию, которая необходима пользователю, осуществляющему поисковый запрос. Скачивание качественных веб-страниц осуществляется за счет применения стратегии отбора, а наличие свежей информации в базе данных поисковой системы осуществляется за счет применения той или иной стратегии повторного обхода. Объединение этих двух стратегий составляет стратегию приоритезации поискового робота.

### 1.4.1. Стратегии отбора

Стратегия отбора определяет, в каком порядке поисковый робот должен обрабатывать веб-страницы, URL-адреса которых были добавлены в очередь на скачивание. Для этого необходимо осуществить эвристическую оценку важности для каждой веб-страницы, являющейся кандидатом на скачивание. В книге [7] стратегии отбора разделены на три категории, описанные ниже.

#### **Стратегии, не использующие дополнительную информацию**

Стратегии, отнесенные в данную категорию, используют информацию, полученную в процессе текущего обхода, и никакую более.

**Обход в ширину.** При использовании данной стратегии поисковый робот скачивает новые страницы, осуществляя обход в ширину веб-графа, начиная с главных страниц стартового множества. То есть веб-страница будет скачана тем раньше, чем раньше ссылка на нее была извлечена с

какой-либо скачанной веб-страницы. В работе [8] показано, что в проведенных экспериментах поисковый робот, использующий данную стратегию, выбирал важные страницы первыми.

**OPIC.** Данный метод описан в работе [1]. Этот метод также базируется на представлении веба как графа, описанному в разделе 1.1.1. В начале обхода каждой странице присваивается одинаковая стоимость. После загрузки страницы ее стоимость поровну распределяется между веб-страницами, на которые она ссылается. Более высокий приоритет имеют страницы с высокой текущей стоимостью. Вычисление стоимости аппроксимирует вычисление *PageRank*.

**Количество обратных ссылок.** Данная стратегия была описана в работе [9]. В этом алгоритме первыми скачиваются страницы, имеющие наибольшее число входящих ссылок.

**FICA.** Метод подробно описан в статье [2]. Алгоритм базируется на методике обучения с подкреплением, где для каждого узла веб-графа вычисляется ранг на основании расстояния между страницами.

## **Стратегии, использующие информацию, полученную ранее**

Стратегии данной категории используют информацию о рангах страниц, полученную в результате предыдущих обходов. При этом поисковый робот начинает обход со страниц, имеющих более высокий ранг, а ранги для страниц, найденных впервые, рассчитываются следующими способами:

- Запрашивая информацию о рангах у «оракула», располагающего полной информацией о веб-графе;
- Равновероятно выбирая одно из значений рангов страниц, полученных в предыдущих обходах;
- Назначаются нулем, то есть сначала скачиваются все страницы, известные ранее, а затем — новые;

- Как значение ранга страницы, указывающей на нее, деленное на число ее исходящих ссылок.

По результатам экспериментов в работе [7], стратегии данной категории преимущественно превосходят стратегии, не использующие дополнительной информации. При этом лучшим является метод, запрашивающий информацию о рангах у «оракула».

### **Стратегия с полной информацией**

К данной категории относится одна стратегия *Omniscient*, использование которой подразумевает наличие «оракула», который вычисляет актуальный ранг для каждой веб-страницы. Для того, чтобы выбрать страницы из очереди на скачивание, поисковый робот осуществляет запрос к «оракулу», после чего выбирает страницы с наибольшим рангом. Данная стратегия быстрее всего находит важные страницы.

#### **1.4.2. Стратегии повторного обхода**

Стратегия повторного обхода определяет какие из уже скачанных страниц следует скачать еще раз. Оба типа стратегий описаны в работе [10].

#### **С одинаковой частотой**

Данная стратегия обхода подразумевает, что веб-страницы будут скачиваться с одинаковой частотой. При этом их важность никак не будет учитываться.

#### **С разными частотами**

При использовании данного типа стратегии, веб-страницы скачиваются с разными частотами. Например, с частотами, пропорциональными частотам изменения страниц.

## Глава 2. Предлагаемый метод

В ходе работы был разработан и реализован программный модуль, реализующий поискового робота. В разделе 2.1 описана предлагаемая стратегия приоритезации, а в разделе 2.2 описаны основные этапы алгоритма работы поискового робота.

### 2.1. СТРАТЕГИЯ ПРИОРИТЕЗАЦИИ

Основная идея стратегии приоритезации заключается в объединении стратегии отбора и стратегии повторного обхода. Это осуществляется посредством введения единого значения  $VRank$  (visit rank). Вычисление такого ранга для веб-страницы  $p$  осуществляется по следующей формуле:

$$VRank(p) = QRank(p) * RVRank(p)$$

где  $QRank(p)$  — эвристическая важность страницы, то есть ранг, характеризующий ее качество (quality rank), а  $RVRank(p)$  — ранг, характеризующий необходимость повторно скачивать веб-страницу  $p$  (re-visit rank).

Вычисление  $QRank$  производится с использованием нейронной сети. Задача нейронной сети — по URL-адресу веб-страницы получить число, имеющее смысл эвристической оценки важности страницы. Для чего на основании URL-адреса строится вектор статических особенностей. Список особенностей представлен в разделе 2.3. Нейронная сеть принимает на вход этот вектор и вычисляет ранг веб-страницы.

Для вычисления  $QRank$  предлагается два метода. Первый базируется исключительно на статических особенностях URL-адреса, и для каждой веб-страницы  $QRank$  вычисляется один раз при добавлении адреса в коллекцию. Второй также использует свойства веб-графа. А именно, после обработки скачанной страницы, ее  $QRank$  распределяется между страницами, ссылки на которые она содержит. Таким образом, при добавлении



нового URL-адреса *QRank* считается при помощи нейронной сети, а затем он может изменяться в процессе дальнейшей работы.

Метод вычисления *RVRank* основан на следующей идее. Заново обходить веб-страницы следует с частотой, пропорциональной частоте их изменения. Для чего для каждой страницы необходимо определить период ее изменения. Это осуществляется посредством итеративного приближения, то есть уточнения значения периода изменения страницы при каждом ее скачивании.

Кроме того, следует учитывать, что существуют динамические страницы, которые генерируются заново при переходе по URL-адресу. Динамические страницы не следует скачивать бесконечно часто. Это осуществляется, во-первых, за счет вероятного небольшого значения *QRank* таких страниц, а во-вторых, путем введения нижней границы для значения периода изменения.

## 2.2. ЭТАПЫ АЛГОРИТМА

Основные компоненты разработанного поискового робота представлены на рисунке 3.1.

Алгоритм проходит включает в себя два этапа: подготовительный и основной. На подготовительном этапе происходит инициализация модуля назначения приоритетов, а на основном — непосредственно обход интернета. Рассмотрим подробнее каждый из этапов.

### Подготовительный этап

На подготовительном этапе происходит обучение нейронной сети, которая входит в модуль назначения приоритетов. В качестве обучающего множества используется набор URL-адресов веб-страниц с известными рангами (например, *PageRank*). Затем в коллекцию добавляется стартовое множество URL-адресов, и для каждого из них вычисляется *QRank*.



Рис. 2.1. Основные программные компоненты поискового робота

## Основной этап

Работа алгоритма на основном этапе происходит циклически. На каждом шаге цикла обрабатывается  $K$  страниц из коллекции, имеющих наибольший  $VRank$ .

Цикл начинается с того, что планировщик запрашивает у менеджера коллекции следующую партию адресов веб-страниц. После чего происходит подсчет  $VRank$  для каждой страницы из коллекции посредством запросов менеджера коллекции к модулю приоритизации. Затем менеджер коллекции формирует партию, содержащую  $K$  адресов веб-страниц, имеющих наибольший  $VRank$  и передает ее планировщику. После чего планировщик последовательно передает URL-адреса для дальнейшей обработки.

Обработка начинается с загрузки веб-страницы, получаемой по текущему URL-адресу. Следует помнить, что при загрузке следует учитывать правила вежливости, описанные в разделе 1.3. И в любой момент времени с каждым хостом должно быть установлено не более одного соединения. После загрузки следует проверить, обновилась ли страница с момента преды-

дущего посещения, чтобы осуществить пересчет предполагаемого периода обновления для текущей страницы.

После чего обработчик содержимого страницы извлекает ссылки с загруженной. После чего из каждой ссылки извлекается URL-адрес. Полученное множество адресов передается менеджеру коллекции. После чего происходит расчет *QRank* для каждого URL-адреса, извлеченного с текущей страницы. Расчет производит модуль приоритезации.

Когда все  $K$  страниц из текущей партии будут обработаны, алгоритм перейдет к следующему шагу цикла.

## 2.3. СТАТИЧЕСКИЕ ОСОБЕННОСТИ URL-АДРЕСА

Для вычисления *QRank* веб-страницы были использованы следующие статические особенности ее URL-адреса. Они были разделены на четыре группы.

### Особенности длины

В эту группу включены особенности, характеризующие длины компонент URL-адреса. Такие как:

- Суммарное число символов URL-адреса;
- Число символов схемы URL-адреса;
- Число символов хоста;
- Число символов пути;
- Число символов запроса.

### Орфографические особенности

В эту группу включены особенности, характеризующие орфографические свойства URL-адреса и его компонент. Такие как:

- Суммарное число цифр, входящих в URL-адрес;

- Суммарное число заглавных букв, входящих в URL-адрес;
- Число цифр, содержащихся в хосте;
- Число цифр, содержащихся в пути URL-адреса;
- Число цифр, содержащихся в запросе URL-адреса;
- Число заглавных букв, содержащихся в пути URL-адреса;
- Число заглавных букв, содержащихся в запросе URL-адреса.

### **Количественные особенности термов**

В данную группу включены особенности, характеризующие количество термов в URL-адресе и его компонентах. Разделение на термы осуществлялось посредством разбиения адреса по символом, не являющимися буквами или цифрами. После чего выделялись следующие особенности:

- Суммарное число термов URL-адреса;
- Число термов хоста;
- Число термов пути;
- Число термов запроса.

### **Словарные особенности термов**

В эту группу включены компоненты, характеризующие частоту реальных слов в компонентах URL-адреса. А именно:

- Частота реальных слов среди термов, входящих в путь URL-адреса;
- Частота реальных слов среди термов, входящих в запрос.

# Глава 3. Экспериментальные результаты

Предлагаемый метод был реализован в виде программной системы на языке *Java* и протестирован.

## 3.1. РЕАЛИЗАЦИЯ

При реализации предлагаемого метода были реализованы следующие программные компоненты:

- Обертка для URL-адреса;
- Менеджер хоста;
- Модуль вежливости;
- Планировщик;
- Загрузчик веб-страниц;
- Обработчик содержимого страницы;
- Компоненты модуля назначения приоритетов:
  - Модуль назначения приоритетов, использующий нейронную сеть;
  - Модуль назначения приоритетов, использующий нейронную сеть и свойства веб-графа;
  - Компонент для извлечения особенностей URL-адреса страницы;
  - Компонент, содержащий нейронную сеть;
  - Словарный модуль;
- Менеджер коллекции.

Основным объектом, с которым происходит работа в программе, является обертка для URL-адреса. Она реализована классом *WebURL*. Каждый экземпляр такого класса содержит мета-информацию о веб-странице,

ее *QRank*, время последнего посещения, предполагаемый период обновления, ссылку на менеджер хоста и т.д.

Менеджер хоста реализован классом *HostController* и предназначен для выполнения правил вежливости для определенного хоста. В каждый момент времени с хостом должно быть установлено только одно соединение. Это реализовано за счет синхронизирования загрузок веб-страницы с одного хоста посредством использования поля *lock* менеджера хоста, являющегося экземпляром класса *Lock*. Кроме того, экземпляр класса хранит на модуль вежливости, который определяет правила доступа к страницам хоста. Также менеджер хранит количество страниц с текущего хоста, имеющих в коллекции.

Модуль вежливости предназначен для соблюдения правил вежливости, описанных в файле *robots.txt*, содержащемся в корневом каталоге многих сайтов. Модуль вежливости реализован классом *PolitenessModule*, содержащий множество (экземпляр класса *TreeSet*) правил, являющихся экземплярами класса *Rule*. Правило хранит текстовый шаблон (*Pattern*), удовлетворяющий набору каталогов сайта, и правило доступа (доступно или нет). Порядок на множестве введен для быстрого определения доступности того или иного URL-адреса.

Планировщик реализован классом *Scheduler*. Планировщик на каждом шаге цикла основного этапа алгоритма формирует  $K$  заданий, являющихся экземплярами исполняемого (*Runnable*) класса *PageProcessingTask*. Эти задания передаются исполнителю (экземпляр класса *ExecutorService*). Задание состоит из следующих инструкций для последовательной обработки. Загрузка, осуществляемая загрузчиком веб-страниц. Обработка содержимого и извлечение ссылок со страницы, осуществляемые обработчиком содержимого страницы. И наконец отправка менеджеру коллекции для дальнейшей обработки. Обработка заданий осуществляется в несколько потоков.

Модуль назначения приоритетов является экземпляром класса, наследующего абстрактный класс *PrioritizationModule*. Разработаны два метода приоритезации. Первый использует только нейронную сеть, а второй — нейронную сеть и анализ веб-графа. Реализованы они соответственно классами *NeuralPrioritizationModule* и *NeuralGraphPrioritizationModule*. Исходный код этих классов приведен в приложениях: А.1 и А.2. Кроме того, были реализованы метод приоритезации *breadth-first* и *FICA* для сравнения с разработанными стратегиями приоритезации.

Компонент для извлечения особенностей URL-адреса реализован классом *FeaturesExtractor*. Исходный программный код класса приведен в приложении А.3. Для каждой из особенностей адреса компонент реализует его численное представление в интервале от 0 до 1 посредством нормировки. Для извлечения словарных особенностей термов конструируется словарный модуль, использующий слова английского и русского языков. Причем слова русского языка были заменены на их транслитерационный вид.

Нейронная сеть была реализована с использованием библиотеки *Encog* [11]. Сеть содержит три слоя нейронов. Входной слой содержит число нейронов, соответствующее числу особенностей URL-адреса веб-страницы (максимальное число особенностей — 16). Слой имеет сигмоидную активирующую функцию. Число нейронов в скрытом слое зависит от числа нейронов входного слоя. А именно — число нейронов скрытого слоя в три раза больше. Скрытый слой так же имеет сигмоидную активирующую функцию. Выходной слой содержит один нейрон и сигмоидную активирующую функцию. Для обучения использовался метод быстрого распространения (*Quickprop*), являющийся улучшением метода обратного распространения ошибки. Подробно он описан в статье [12]. Данная конфигурация сети была подобрана в ходе тестирования на множестве URL-адресов веб-страниц с известными *PageRank*, полученными с использованием *google toolbar*.

Именно с такими параметрами отклонение вычисленного ранга от реального *PageRank* было минимальным.

Менеджер коллекции реализован классом *CollectionHandler*. Он содержит ссылку на модуль приоритезации, посредством которого вычисляются ранги для URL-адресов веб-страниц, передаваемых в коллекцию. Коллекция представлена экземпляром класса *ConcurrentHashMap*, где ключом является строка, а значением — *WebURL*. Выбрана потокобезопасная реализация ассоциативного массива, поскольку добавление адресов происходит в несколько потоков. Также менеджер коллекции содержит множество менеджеров хостов для того, чтобы контролировать количество добавляемых в коллекцию страниц. Кроме того, менеджер коллекции осуществляет формирование пакета веб-страниц для обработки по запросу планировщика. Формирование осуществляется путем подсчета *VRank* для каждой страницы коллекции и выбора *K* страниц с наибольшим таким рангом. Для эффективной реализации такой сортировки был реализован компаратор *ValueComparator*, осуществляющий сортировку в *Map* по значению. В процессе формирования партии веб-страниц контролируется, чтобы она не содержала более определенного в планировщике числа страниц с одного сайта. Это необходимо для того, чтобы поисковый робот не заикливался на одном сайте, что привело бы к замедлению работы.

### 3.2. ПАРАМЕТРЫ ЗАПУСКОВ

Поисковый робот тестировался на сайтах, принадлежащих домену *.ru*. Обработка страниц происходила в 20 параллельных потоков.

Нейронная сеть обучалась на случайной выборке размером 50000 URL-адресов с домена *.ru*, для которых были известны их *PageRank*. Ранги были получены посредством использования *google toolbar*.

Максимальное число обрабатываемых сайтов составляло 5000. Раз-



мер партии страниц, обрабатываемых на одной итерации цикла составляло  $K = 1000$ . Максимальное число страниц с одного сайта в пакете составляло 50.

### 3.3. РЕЗУЛЬТАТЫ

Для сравнения работы различных стратегий приоритезации были произведены тестовые запуски поискового робота. Стартовое множество URL-адресов состояло из главных страниц популярных сайтов, принадлежащих домену *.ru*. Размер стартового множества — 100 сайтов. В процессе обхода интернета при каждом запуске в коллекцию добавлялось около 200 тысяч URL-адресов, из которых были скачаны 25 тысяч страниц с адресами, имеющими наибольший приоритет.

Были протестированы стратегии приоритезации, реализующие два предложенных метода, а также стратегия, использующая обход в ширину. Для всех скачанных в процессе запусков страниц был получен их *PageRank* с использованием сервиса *googletoolbar*, являющийся показателем их качества. На рисунке ?? приведен график зависимости среднего *PageRank* страниц от числа скачанных для каждой из протестированных стратегий приоритезации.

Таким образом, оба разработанных метода превосходят стандартный метод обхода в ширину, поскольку *PageRank* скачанных страниц поисковым роботом, использующим любую из разработанных стратегий приоритезации, превосходит *PageRank* скачанных страниц поисковым роботом, использующим обход в ширину.

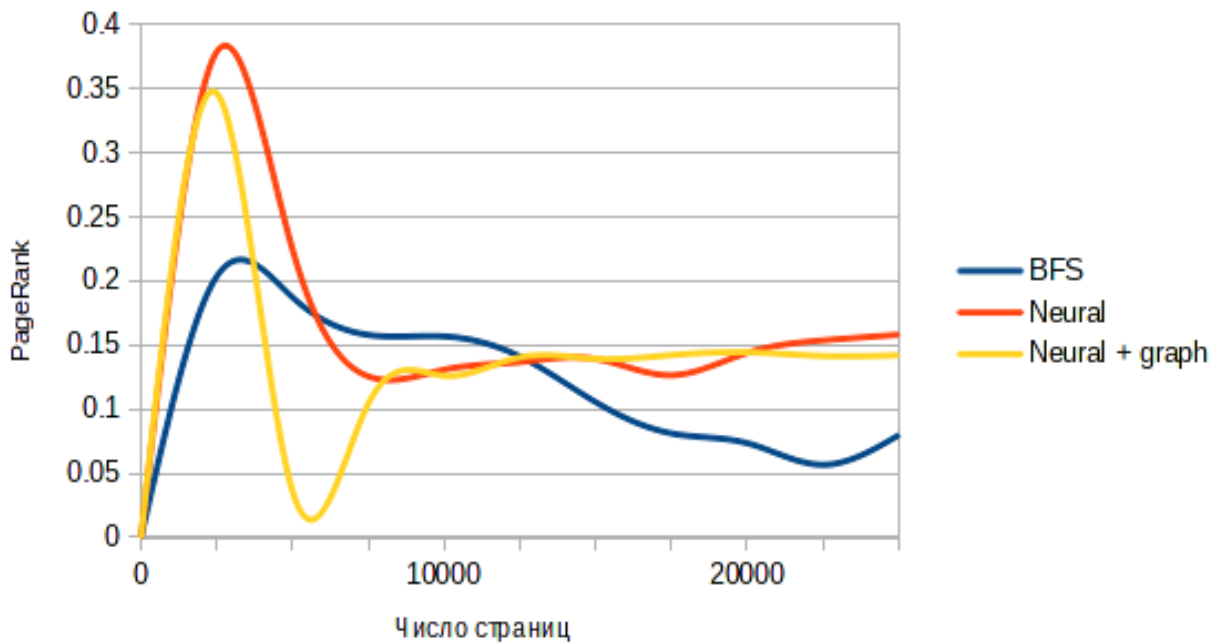


Рис. 3.1. График зависимости среднего *PageRank* страниц от скачанного числа страниц. *BFS* — стратегия обхода в ширину. *Neural* — стратегия приоритезации, использующая нейронную сеть. *Neural + Graph* — стратегия приоритезации использующая нейронную сеть и веб-граф

## Заключение

В работе получены следующие результаты. Предложены и разработаны две стратегии приоритезации поискового робота, использующие статические признаки URL-адресов.

Тестовые запуски показали, что поисковый робот, использующий любую из них, скачивает более качественные страницы, чем поисковый робот, использующий обход в ширину.

# Источники

- [1] *Abiteboul S., Preda M., Cobena G.* Adaptive on-line page importance computation // Proceedings of the 12th international conference on World Wide Web. WWW '03. New York, NY, USA: ACM, 2003. Pp. 280–290. <http://doi.acm.org/10.1145/775152.775192>.
- [2] *Bidoki A. M. Z., Yazdani N., Ghodsnia P.* Fica: A novel intelligent crawling algorithm based on reinforcement learning // *Web Intelli. and Agent Sys.* 2009. Vol. 7, no. 4. Pp. 363–373. <http://dx.doi.org/10.3233/WIA-2009-0174>.
- [3] *Richardson M., Prakash A., Brill E.* Beyond pagerank: machine learning for static ranking // Proceedings of the 15th international conference on World Wide Web. WWW '06. New York, NY, USA: ACM, 2006. Pp. 707–715. <http://doi.acm.org/10.1145/1135777.1135881>.
- [4] *Brin S., Page L.* The anatomy of a large-scale hypertextual web search engine // *Comput. Netw. ISDN Syst.* 1998. Vol. 30, no. 1-7. Pp. 107–117. [http://dx.doi.org/10.1016/S0169-7552\(98\)00110-X](http://dx.doi.org/10.1016/S0169-7552(98)00110-X).
- [5] *Wikipedia.* Алгоритм HITS — Wikipedia, The Free Encyclopedia. 2013. [Online; accessed 15-June-2013].
- [6] *Manning C. D., Raghavan P., Schütze H.* Introduction to Information Retrieval. New York, NY, USA: Cambridge University Press, 2008.
- [7] *Castillo C.* Effective web crawling // *SIGIR Forum.* 2005. Vol. 39, no. 1. Pp. 55–56. <http://doi.acm.org/10.1145/1067268.1067287>.
- [8] *Najork M., Wiener J. L.* Breadth-first crawling yields high-quality pages // Proceedings of the 10th international conference on World Wide Web. WWW '01. New York, NY, USA: ACM, 2001. Pp. 114–118. <http://doi.acm.org/10.1145/371920.371965>.
- [9] *Cho J., Garcia-Molina H., Page L.* Efficient crawling through url ordering // *Comput. Netw. ISDN Syst.* 1998. Vol. 30, no. 1-7. Pp. 161–172. [http://dx.doi.org/10.1016/S0169-7552\(98\)00108-1](http://dx.doi.org/10.1016/S0169-7552(98)00108-1).
- [10] *Cho J., Garcia-Molina H.* The evolution of the web and implications for an incremental crawler // Proceedings of the 26th International Conference on Very Large Data Bases. VLDB '00. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2000. Pp. 200–209. <http://dl.acm.org/citation.cfm?id=645926.671679>.
- [11] Encog Machine Learning Framework. <http://www.heatonresearch.com/encog>.
- [12] *Fahlman S. E.* An empirical study of learning speed in back-propagation networks: Tech. rep.: 1988.

# Приложение А

## Программный код

### А.1. ПРОГРАММНЫЙ КОД МОДУЛЯ МОДУЛЯ НАЗНАЧЕНИЯ ПРИОРИТЕТОВ, ИСПОЛЬЗУЮЩЕГО НЕЙРОННУЮ СЕТЬ

```
1 package ru.ifmo.mailru.priority;

3 import ru.ifmo.mailru.core.Page;
  import ru.ifmo.mailru.core.WebURL;
5 import ru.ifmo.mailru.features.NeuralPrioritizationController;

7 import java.io.FileNotFoundException;

9 /**
   * @author Anastasia Lebedeva
11 */
12 public class NeuralPrioritizationModule extends PrioritizationModule {
13     private NeuralPrioritizationController neuralPrioritizationController
        ;

15     public NeuralPrioritizationModule() throws FileNotFoundException {
        neuralPrioritizationController = new
            NeuralPrioritizationController();
17         int positiveSize = neuralPrioritizationController.
            positiveExamples.size();
        int negativeSize = neuralPrioritizationController.
            negativeExamples.size();
19         neuralPrioritizationController.train(positiveSize, negativeSize);
20     }

21     @Override
22     public void setQualityRanks(WebURL url, Page parentPage) {
        Double computedRank = neuralPrioritizationController.computeRank(
            url);
24         url.setQualityRank(computedRank);
25     }

26     @Override
27     public void resetQualityRanks(WebURL url, Page parentPage) {
        //To change body of implemented methods use File | Settings |
        File Templates.
29     }

31 }
```

```
}
```

Листинг А.1. Модуль назначения приоритетов, использующий нейронную сеть

## А.2. ПРОГРАММНЫЙ КОД МОДУЛЯ МОДУЛЯ НАЗНАЧЕНИЯ ПРИОРИТЕТОВ, ИСПОЛЬЗУЮЩЕГО НЕЙРОННУЮ СЕТЬ И СВОЙСТВА ВЕБ-ГРАФА

```
package ru.ifmo.mailru.priority;
2
import ru.ifmo.mailru.core.Page;
4 import ru.ifmo.mailru.core.WebURL;
import ru.ifmo.mailru.features.NeuralPrioritizationController;
6
import java.io.FileNotFoundException;
8
/**
10 * @author Anastasia Lebedeva
*/
12 public class NeuralGraphPrioritizationModule extends PrioritizationModule
{
    private NeuralPrioritizationController neuralPrioritizationController
        ;
14
    public NeuralGraphPrioritizationModule() throws FileNotFoundException
    {
16        neuralPrioritizationController = new
            NeuralPrioritizationController();
        int positiveSize = neuralPrioritizationController.
            positiveExamples.size();
18        int negativeSize = neuralPrioritizationController.
            negativeExamples.size();
        neuralPrioritizationController.train(positiveSize, negativeSize);
20    }

22
    @Override
24    public void setQualityRanks(WebURL url, Page parentPage) {
        double addition = 0;
26        if (parentPage != null) {
            addition = parentPage.getUrl().getQualityRank() / parentPage.
                getOutLinks().size();
28        }
        Double computedRank = neuralPrioritizationController.computeRank(
            url);
30        url.incrementQualityRank(computedRank + addition);
    }

32
    @Override
34    public void resetQualityRanks(WebURL url, Page parentPage) {
        double addition = parentPage.getUrl().getQualityRank() /
            parentPage.getOutLinks().size();
36        url.incrementQualityRank(addition);
    }
38 }
```

Листинг А.2. Модуль назначения приоритетов, использующий нейронную сеть и свойства веб-графа

### А.3. ПРОГРАММНЫЙ КОД МОДУЛЯ ВЫДЕЛЕНИЯ СТАТИЧЕСКИХ ОСОБЕННОСТЕЙ URL-АДРЕСА

```
package ru.ifmo.mailru.features;

import ru.ifmo.mailru.core.WebURL;

import java.net.URI;
import java.net.URISyntaxException;
import java.util.ArrayList;
import java.util.List;
import java.util.regex.Matcher;
import java.util.regex.Pattern;

public class FeaturesExtractor {
    private List<Double> features;
    private String[] components;
    private final int COMPONENT_NUMBER = 5;
    private static double[][] MAXIMUMS;

    /**
     * row 0: length features
     * row 1: orthographic features (digits)
     * row 2: orthographic features (capital case)
     * row 3: number of term
     * row 4: dictionary features
     */

    {
        MAXIMUMS = new double[COMPONENT_NUMBER][4];
        for (int j = 0; j < 3; j++) {
            for (int i = 3; i < COMPONENT_NUMBER; i++) {
                MAXIMUMS[i][j] = 300;
            }
        }
        for (int i = 0; i < 3; i++) {
            MAXIMUMS[0][i] = 650;
            MAXIMUMS[1][i] = 5;
            MAXIMUMS[2][i] = 50;
        }
        for (int i = 3; i < COMPONENT_NUMBER; i++) {
            MAXIMUMS[i][3] = 100;
        }
        MAXIMUMS[0][3] = 300;
        MAXIMUMS[1][3] = 1;
        MAXIMUMS[2][3] = 10;
    }

    public FeaturesExtractor(String url) throws URISyntaxException {
        this(new WebURL(new URI(url)));
    }

    public FeaturesExtractor(WebURL url) {
        components = new String[COMPONENT_NUMBER];
        URI uri = url.getUri();
        components[0] = uri.toString();
        components[1] = uri.getScheme();
        components[2] = uri.getHost();
    }
}
```

```

56     components[3] = uri.getPath();
57     components[4] = uri.getQuery();
58     for (int i = 1; i < COMPONENT_NUMBER; i++) {
59         if (components[i] == null) {
60             components[i] = "";
61         }
62     }
63     features = new ArrayList<>();
64 }

65 private void extractLengthFeatures() {
66     for (int i = 0; i < COMPONENT_NUMBER; i++) {
67         features.add(components[i].length() / MAXIMUMS[i][0]);
68     }
69     String[] path = components[3].split("/");
70     features.add((double) path.length / 20);
71 }

72 private void extractOrthographicFeatures() {
73     Pattern digit = Pattern.compile("\\d+");
74     double commonCount = 0;
75     for (int i = 2; i < COMPONENT_NUMBER; i++) {
76         Matcher m = digit.matcher(components[i]);
77         double count = 0;
78         while (m.find()) {
79             count += m.group().length();
80         }
81         commonCount += count;
82         features.add(count / MAXIMUMS[i][1]);
83     }
84     features.add(commonCount / MAXIMUMS[0][1]);
85     Pattern capitalCase = Pattern.compile("[A-Z]+");
86     commonCount = 0;
87     for (int i = 3; i < COMPONENT_NUMBER; i++) {
88         Matcher m = capitalCase.matcher(components[i]);
89         double count = 0;
90         while (m.find()) {
91             count += m.group().length();
92         }
93         commonCount += count;
94         features.add(count / MAXIMUMS[i][2]);
95     }
96     features.add(commonCount / MAXIMUMS[0][2]);
97 }

98 private void extractTermFeature() {
99     List<String> terms = extractTerms();
100     extractCountTerm(terms);
101     extractWordFrequency(terms);
102 }

103 private void extractWordFrequency(List<String> terms) {
104     for (int i = 3; i < COMPONENT_NUMBER; i++) {
105         if (terms.get(i).length == 0) {
106             features.add(1.0);
107             continue;
108         }
109         double words = 0;
110         for (String s : terms.get(i)) {
111             if (DictionaryModule.isWord(s)) {

```

```

116         words++;
117     }
118 }
119     features.add(words / terms.get(i).length);
120 }
121 }
122
123 private void extractCountTerm(List<String[]> terms) {
124     for (int i = 0; i < COMPONENT_NUMBER; i++) {
125         features.add(terms.get(i).length / MAXIMUMS[i][3]);
126     }
127 }
128
129 private List<String[]> extractTerms() {
130     List<String[]> terms = new ArrayList<>(COMPONENT_NUMBER);
131     for (int i = 0; i < COMPONENT_NUMBER; i++) {
132         terms.add(DictionaryModule.splitIntoTokens(components[i]));
133     }
134     return terms;
135 }
136
137
138 public Double[] buildVector() {
139     extractLengthFeatures();
140     extractOrthographicFeatures();
141     extractTermFeature();
142     Double[] res = new Double[features.size()];
143     return features.toArray(res);
144 }
145 }

```

Листинг А.3. Модуль выделения статических особенностей URL-адреса