

# Глава 1

## Алгоритмы

### 1.1 Классы данных

**TransitionMatrix** Класс представляет собой обертку над матрицей. В данной матрице  $M$ :

$M[i][j] = 0$ , если не существует прямой ссылки со страницы  $i$  на страницу  $j$ ,

$M[i][j] = 1$ , если ссылка со страницы  $i$  на страницу  $j$  существует.

**TransitionProbabilityMatrix** Класс представляет собой обертку над матрицей. В данной матрице  $M$ :

$M[i][i] = d$ ,

$M[i][j] = p_{i,j} * (1 - d), i \neq j$ , где  $d$  - коэффициент затухания или

вероятность пользователя остаться на странице,  $1 - d$  - вероятность перейти на другую страницу,  $p_{i,j}$  - вероятность перехода с  $i$  страницы на  $j$ .

**TransitionList** Класс представляет собой набор списков ссылок для каждой страницы, то есть:

Для  $i$ -го списка наличие элемента со значением  $j$  означает наличие ссылки с  $i$  страницы в  $j$ .

## 1.2 Интерфейс алгоритмов

**Запуск** Все алгоритмы реализованы классами и описаны в файле `ComputingMethod.py`. Все объекты инициализируются некоторыми стартовыми параметрами, после чего их можно использовать следующими путями:

- 1) *iterating\_run*( $N$ ) - запуск  $N$  итераций выбранного алгоритма
- 2) *stopping\_run*( $\epsilon$ ) - запуск алгоритма до срабатывания критерия остановки, где  $\epsilon$  - заданная точность вычислений

**Замечания** Не все классы определяют сразу оба пути использования алгоритмов, некоторые из них содержат только один из них. При использовании нереализованного функционала происходит оста-

новка программы с *Exception("Noavailableimplementation")*.

Все классы опираются на метод *\_iteration*, осуществляющий одну итерацию алгоритма. Его использование возможно, но приведет к потере функционала либо производительности данного метода.

Все алгоритмы поддерживают изменение коэффициента затухания  $d$  теми или иными способами, но также имеют значение по умолчанию  $d = 0.85$ , так что задавать его вручную не обязательно.

Все итеративные алгоритмы по умолчанию определяют некоторое значение точности  $\epsilon$ , все представленные тесты были запущены с базовыми значениями параметров.

**Данные** Классы содержат три ключевых поля:

*\_iterating\_run\_time* - время выполнения последнего *iterating\_run*,  
None по умолчанию

*\_stopping\_run\_time* - время выполнения последнего *stopping\_run*,  
None по умолчанию

*\_stat\_vector* - вычисленный вектор, формат вектора определяется алгоритмом, как и его значение по умолчанию

## 1.3 Итеративные алгоритмы

### Цепь Маркова

**Описание** Имеется матрица переходных вероятностей Марковской цепи  $A$ , где  $A[i][j], i \neq j$  - вероятность перехода со страницы  $i$  на страницу  $j$ .  $A[i][i] = 0$ .  $d$  - коэффициент затухания или вероятность пользователя остаться на странице.  $1 - d$  есть вероятность перейти на другую страницу. Матрица  $H$  есть преобразованная следующим образом матрица  $A$ :

$$H[i][i] = d,$$

$$H[i][j] = A[i][j] * (1 - d), i \neq j$$

Данный алгоритм итеративно ищет приближенное значение вектора статистических вероятностей. Критерием остановки является достаточная близость статистических векторов на текущей и предыдущей итерации или:

$$|\pi_k - \pi_{k-1}| < \epsilon, \text{ где } \epsilon - \text{заранее заданная точность вычислений.}$$

**Реализация** Алгоритм реализован классом `MarkovChain`. При создании объекта класса необходимо передать `TransitionProbabilityMatrix`. Необязательным параметром является стартовый статистический вектор, который по умолчанию задается как вектор одинаковых вероятностей  $1/N$ , где  $N$  - количество страниц.

Реализация поддерживает оба формата запуска.

## Power Method

**Описание** Имеется матрица переходов  $A$ , где:

$[i][j] = 0$ , если не существует прямой ссылки со страницы  $i$  на страницу  $j$ ,

$[i][j] = 1$ , если ссылка со страницы  $i$  на страницу  $j$  существует.

Каждая страница имеет некоторый вес, при этом:

1) Собственный вес страницы  $W$  есть сумма всех входящих весов ссылающихся страниц с коэффициентом  $d$  плюс константа  $1 - d$ :

$$W = d * \sum W_{in} + 1 - d$$

2) Исходящий вес страницы есть  $W/N$ , где  $W$  - собственный вес страницы,  $N$  - количество ссылок.

Алгоритм итеративно ищет приближенное значение вектора весов страниц. Критерием остановки является достаточная близость векторов весов на текущей и предыдущей итерации или:

$$|\pi_k - \pi_{k-1}| < \epsilon, \text{ где } \epsilon - \text{заранее заданная точность вычислений.}$$

**Реализация** Алгоритм реализован классом `PowerMethod`. При создании объекта класса необходимо передать `TransitionMatrix`. Необязательным параметром является стартовый вектор весов, который по умолчанию задается как вектор единиц. Коэффициент затухания также может быть передан.

Реализация поддерживает оба формата запуска.

## Adaptive Power Method

**Описание** Алгоритм отличается от предыдущего лишь критерием остановки: пересчет останавливается не для всего вектора сразу, а для отдельных компонент, когда разница между значениями достаточно маленькой или:

$$|\pi_k^i - \pi_{k-1}^i| < \epsilon, \text{ где } \epsilon - \text{заранее заданная точность вычислений.}$$

Как только пересчет всех компонент вектора был остановлен, алгоритм заканчивает выполняться.

**Реализация** Алгоритм реализован классом `AdaptivePowerMethod`. При создании объекта класса необходимо передать `TransitionMatrix`. Необязательным параметром является стартовый вектор весов, который по умолчанию задается как вектор единиц. Коэффициент затухания также может быть передан.

Реализация поддерживает оба формата запуска.

## Extrapotaling Adaptive Power Method

**Описание** Алгоритм является улучшением предыдущего. Раз в  $L$  шагов алгоритм высчитывает уточняющий коэффициент:  $\lambda_2^k \gamma_2 y_2^T = \frac{(\pi^{(k+1)T} - \pi^{(k)T})^2}{\pi^{(k+2)T} - 2\pi^{(k+1)T} - \pi^{(k)T}}$ . Вычисленный коэффициент вычитается из  $\pi^{(k)T}$ , давая улучшенное приближение вектора весов.

**Реализация** Алгоритм реализован классом `ExtrapolatingAdaptivePowerMethod`. При создании объекта класса необходимо передать `TransitionMatrix`. Необязательным параметром является стартовый вектор весов, который по умолчанию задается как вектор едениц. Коэффициент затухания также может быть передан.

Реализация поддерживает оба формата запуска, при этом можно задать период экстраполяции -  $L$  (10 по умолчанию, должен быть больше 3-х).

## 1.4 Алгоритмы Монте Карло

### Endpoint Random start Monte Carlo Method

**Описание** Задан список ссылок  $S$ , где  $S[i]$  - список страниц, на которые есть ссылки с страницы  $i$ , то есть:

$if j \in S[i]$ , то есть ссылка с  $i$  страницы на страницу  $j$ .

Симуляция прохода - симуляция перемещения пользователя по страницам, при этом:

1) Переход пользователя на любую страницу, на которую имеется ссылка, равновероятен.

2) Переход пользователя по страницам заканчивается с вероятностью  $1 - d$

Метод выполняет  $n$  или  $n^2$  симуляций проходов, *начинающихся со случайной страницы*, за одну итерацию. Страница, на которой симуляция заканчивается увеличивает свой вес на 1.

**Реализация** Алгоритм реализован классом EndpointRandomStartMonteCarloM. При создании объекта класса необходимо передать TransitionList. Коэффициент затухания может быть передан.

Реализация поддерживает только *iterating\_run*, предоставляя на выбор два возможных количества прогонов за итерацию:

linear -  $n$ ,

square -  $n^2$ , где  $n$  - количество страниц.

**Оценка сходимости**  $E(\hat{\pi}_j, N) = \pi_j$

$Var(\hat{\pi}_j, N) = N^{-1}\pi_j(1 - \pi_j) < 1/(4N)$

## Endpoint Cyclic Start Monte Carlo Method

**Описание** Задан список ссылок  $S$ , где  $S[i]$  - список страниц, на которые есть ссылки с страницы  $i$ , то есть:

*if*  $j \in S[i]$ , то есть ссылка с  $i$  страницы на страницу  $j$ .

Симуляция прохода - симуляция перемещения пользователя по страницам, при этом:



1) Переход пользователя на любую страницу, на которую имеется ссылка, равновероятен.

2) Переход пользователя по страницам заканчивается с вероятностью  $1 - d$

Метод выполняет *заданное количество симуляций проходов для каждой страницы* за одну итерацию. Страница, на которой симуляция заканчивается увеличивает свой вес на 1.

**Реализация** Алгоритм реализован классом EndpointCyclicStartMonteCarloMe. При создании объекта класса необходимо передать TransitionList. Коэффициент затухания может быть передан.

Реализация поддерживает только *iterating\_run*, предоставляя возможность выбора количества запусков симуляций прохода с каждой страницы - m (3 по умолчанию).

**Оценка сходимости**  $E(\hat{\pi}_j) = \pi_j$   
 $Var(\hat{\pi}_j) = N^{-1}[\pi_j - n^{-1} \sum_{i=1}^n p_{ij}^2] < Var(\hat{\pi}_j, N)$

## Complete Path Monte Carlo Method

**Описание** Задан список ссылок S, где  $S[i]$  - список страниц, на которые есть ссылки с страницы i, то есть:

*if*  $j \in S[i]$ , то есть ссылка с  $i$  страницы на страницу  $j$ .

Симуляция прохода - симуляция перемещения пользователя по страницам, при этом:

- 1) Переход пользователя на любую страницу, на которую имеется ссылка, равновероятен.
- 2) Переход пользователя по страницам заканчивается с вероятностью  $1 - d$

Метод выполняет заданное количество симуляций проходов для каждой страницы за одну итерацию. *Каждая посещенная в процессе симуляции страница* увеличивает свой вес на 1.

**Реализация** Алгоритм реализован классом CompletePathMonteCarloMethod. При создании объекта класса необходимо передать TransitionList. Коэффициент затухания может быть передан.

Реализация поддерживает только *iterating\_run*, предоставляя возможность выбора количества запусков симуляций прохода с каждой страницы -  $m$  (3 по умолчанию).

## Stopping Complete Path Monte Carlo Method

**Описание** Задан список ссылок  $S$ , где  $S[i]$  - список страниц, на которые есть ссылки с страницы  $i$ , то есть:

*if*  $j \in S[i]$ , то есть ссылка с  $i$  страницы на страницу  $j$ .

Симуляция прохода - симуляция перемещения пользователя по страницам, при этом:

1) Переход пользователя на любую страницу, на которую имеется ссылка, равновероятен.

2) Переход пользователя по страницам заканчивается с вероятностью  $1 - d$  или *при попадании на страницу без внешних ссылок*.

Метод выполняет заданное количество симуляций проходов для каждой страницы за одну итерацию. Каждая посещенная в процессе симуляции страница увеличивает свой вес на 1.

**Реализация** Алгоритм реализован классом `StoppingCompletePathMonteCarlo`. При создании объекта класса необходимо передать `TransitionList`. Коэффициент затухания может быть передан.

Реализация поддерживает только *iterating\_run*, предоставляя возможность выбора количества запусков симуляций прохода с каждой страницы -  $m$  (3 по умолчанию).

Метод выполняет заданное число симуляций проходов для каждой страницы, завершающихся с вероятностью  $(1 - d)$  или при попадании на страницу без внешних ссылок. Все страницы, посещенные в процессе прохода, увеличивают показатель ранга на 1. Возвращает итоговый вектор сумм.

## Random start Stopping Complete Path Monte Carlo Method

**Описание** Задан список ссылок  $S$ , где  $S[i]$  - список страниц, на которые есть ссылки с страницы  $i$ , то есть:

*if  $j \in S[i]$ , то есть ссылка с  $i$  страницы на страницу  $j$ .*

Симуляция прохода - симуляция перемещения пользователя по страницам, при этом:

1) Переход пользователя на любую страницу, на которую имеется ссылка, равновероятен.

2) Переход пользователя по страницам заканчивается с вероятностью  $1 - d$  или *при попадании на страницу без внешних ссылок.*

Метод выполняет  $n$  или  $n^2$  симуляций проходов, *начинающихся со случайной страницы*, за одну итерацию. Каждая посещенная в процессе симуляции страница увеличивает свой вес на 1.

**Реализация** Алгоритм реализован классом RandomStartStoppingCompletePat. При создании объекта класса необходимо передать TransitionList. Коэффициент затухания может быть передан.

Реализация поддерживает только *iterating\_run*, предоставляя на выбор два возможных количества прогонов за итерацию:

linear -  $n$ ,

square -  $n^2$ , где  $n$  - количество страниц.

## Глава 2

# Тесты

### 2.1 Описание тестов

Все тесты (кроме Level Test) используют заранее подготовленные файлы с известными ответами. Первый и третий файл - примеры, найденные в интернете. Второй был сгенерирован с помощью SeparatedLevelsTreeGen, результат на нем просчитан с помощью Power Method.

#### Position Test

**Описание** Тест создает списки страниц, упорядоченные по убыванию ранга, для истинного и полученного результатов. После этого считает количество совпадений элементов, расположенных на местах

с одинаковым номером и возвращает процент совпадений:

$$\text{Position Test} = \frac{\sum_{i=0}^N I(r_i=t_i)}{N}, \text{ где}$$

$N$  - количество страниц,

$I$  - индикатор выражения, возвращает 1, если выражение истинно, и 0, если выражение ложно,

$r_i$  -  $i$ -я по рангу страница, изначальный список ранжирования есть результат работы алгоритма.

$t_i$  -  $i$ -я по рангу страница, изначальный список есть истинный список ранжирования.

Тест принимает значение от 0 до 1, где 1 - абсолютное совпадение, 0 - абсолютное несовпадение.

**Оценка метрики** Тест оценивает простой и понятный параметр, который легко интерпретировать. Но простота этого теста также его главный недостаток. Ряд ситуаций могут быть неправильно интерпретированы, если рассматривать их со стороны этого теста, например:

1) Когда одна или несколько страниц оказались не на своей позиции, все сдвинутые страницы будут расценены как ошибочно определенные.

2) Алгоритм, который хорошо определяет страницы с наибольшими рангами, будет иметь меньшее значение теста чем алгоритм,

который работает лучше в среднем, хотя определение первых страниц является более важной задачей.

## Sequence Test

**Описание** Тест создает списки страниц, упорядоченные по убыванию ранга, для истинного и полученного результатов. После этого считает количество совпадений элементов, расположенных на местах с одинаковым номером. Если номера страниц не совпадают, то страница, стоящая в списке полученного результата, удаляется.

$\text{Sequence Test} = \frac{D}{N}$ , где

$N$  - количество страниц,

$D$  - количество страниц, после удаления всех несовпадающих страниц.

Тест принимает значение от 0 до 1, где 1 - абсолютное совпадение, 0 - абсолютное несовпадение.

**Оценка метрики** Тест является улучшением предыдущего. Данные изменения позволяют получить более правильную оценку в 1 ситуации предшествующего теста. Так же тест сглаживает неточности и проверяет более процент сопадения последовательностей чем процент полного совпадения. Но при этом:

1) Метрика не дает серьезных улучшений для случая 2 предыдущего алгоритма.

2) Метрика слабо отражает выбросы. Если страница с наименьшим рангом алгоритмом определится первой, при этом  $N = 10000$  и все остальные страницы определятся правильно, то ошибка составит всего  $1/N = 0.0001$ .

## Vector Test

**Описание** Тест нормализует истинный и полученный списки рангов, после чего возвращает расстояние между векторами по метрике l1.

$\text{Vector Test} = \sum_{i=0}^N |R_i - T_i|$ , где

$N$  - количество страниц,

$R_i$  -  $i$ -я по рангу страница, изначальный список ранжирования есть результат работы алгоритма.

$T_i$  -  $i$ -я по рангу страница, изначальный список есть истинный список ранжирования.

Тест принимает значение 0, когда вектора полностью совпали, и 2, когда вектора максимальное расстояние по l1.

**Замечание** Так как вектора нормализованны, то



$\sum_{i=0}^N R_i = 1, \sum_{i=0}^N T_i = 1$ , тогда:

1) если для любой пары  $(R_i, T_i)$  выполняется  $R_i = 0$  or  $T_i = 0$ , то  $\sum_{i=0}^N |R_i - T_i| = \sum |R_k| + \sum |T_j| = 1 + 1 = 2$  где  $i \neq j$

2) если для пар из множества  $Z$  это свойство не выполнено, то для всех пар со свойством их сумма  $\sum |Rz_i - Tz_i| = \sum |Rz_k| + \sum |Tz_j|$ .  $\sum_{z \in Z} |R_z - T_z| = c - \delta$ , где  $c$  - сумма всех компонент пар  $z \in Z$ , а  $\delta$  - сумма минимумов для каждой пары  $z \in Z$ . Тогда:

$$\begin{aligned} \sum_{i=0}^N |R_i - T_i| &= \sum |Rz_i - Tz_i| + \sum_{z \in Z} |R_z - T_z| = \sum |Rz_k| + \\ &+ \sum |Tz_j| + c - \delta = (\sum |Rz_k| + \sum |Tz_j| + c) - \delta = \sum |R_k| + \sum |T_j| - \delta < \\ &\sum |R_k| + \sum |T_j| = 2 \end{aligned}$$

Значит, 2 - максимальное значение данного теста.

**Оценка метрики** Метрика оценивает покомпонентное расстояние между векторами. Данное значение просто интерпретировать, метрика хорошо определяет некоторые выбросы, например, когда у какой-то страницы сильно завышен или занижен ранг, а также ситуации, когда ранги достаточно сильно схожи. При этом метрика слабо выявляет ошибки первых страниц, например, если вектора в среднем похожи.

## Distance Test

**Описание** Тест считает среднее расстояние страниц от их истинного положения, то есть:

$$\text{Distance Test} = \frac{\sum_{i=0}^N |RP[i] - TP[i]|}{N}, \text{ где}$$

$N$  - количество страниц,

$RP[i]$  - место  $i$ -той страницы в списке ранжирования алгоритма.

$TP[i]$  - место  $i$ -той страницы в истинном списке ранжирования.

Тест принимает значение 0, когда вектора полностью совпали. Любое значение больше означает наличие различий.

**Оценка метрики** Метрика служит для определения степени ошибки алгоритма в среднем и работает больше как дополнение к Sequence Test, позволяя лучше анализировать и понимать результаты. Как самостоятельная метрика имеет множество недостатков: зависимость от выбросов, качество не зависит от места ошибок, зависимость от сдвига страниц.

## Level Test

**Описание** Тест использует BinaryTreeGenerator, для генерации графа, который являлся бы бинарным деревом. Для данного дерева нам наперёд известны позиции страниц в ранговом списке. Тест возвращает процент страниц с правильно определённым уровнем.

$\text{Level Test} = \frac{\sum_{i=0}^N I(RL[i]=TL[i])}{N}$ , где

$N$  - количество страниц,

$I$  - индикатор выражения, возвращает 1, если выражение истинно, и 0, если выражение ложно,

$RL[i]$  - уровень  $i$ -й по рангу страницы, изначальный список ранжирования есть результат работы алгоритма.

$TL[i]$  - уровень  $i$ -й по рангу страницы, изначальный список есть истинный список ранжирования.

Тест принимает значение от 0 до 1, где 1 - абсолютное совпадение, 0 - абсолютное несовпадение.

**Оценка метрики** Тест является по сути оценкой разумности для всех алгоритмов, чей результат не является случайным. Так как любое значение кроме 1 будет означать наличие ошибки в работе алгоритма.

Для случайных алгоритмов является показателем работы с простым случаем, а так же неплохой проверкой того, как алгоритм оценивает страницы с одинаковыми рангами.

## Top Test

**Описание** Тест считает процент совпадений первых  $i$  страниц, для всех  $i$  меньших заданного числа, и возвращает список процентов сопадений для каждого  $i$ , то есть:

$$\text{Top Test } [j] = \frac{\sum_{i=0}^j I(RS[0:j][i] \in TS[0:j])}{j}, \text{ где}$$

$j$  - количество первых страниц,

$I$  - индикатор выражения, возвращает 1, если выражение истинно, и 0, если выражение ложно,

$RS[0 : j][i]$  -  $i$ -я страница в срезе, изначальный список ранжирования есть результат работы алгоритма.

$TS[0 : j]$  - множество первых  $j$  страниц в истинном списке ранжирования.

$j$ -я ячейка теста принимает значение от 0 до 1, где 1 - абсолютное совпадение множеств, 0 - абсолютное несовпадение.

**Оценка метрики** Метрика является отличным параметром качества определения топа страниц, то есть страниц, имеющих наибольшие ранги. Это очень важный параметр, так как пользователь чаще всего переходит по странице, находящейся в первой пятёрке. При этом тест не может быть использован отдельно от остальных метрик, так как совсем не оценивает правильность определения рангов на для остальных страниц.

Порог страниц данного теста не следует брать слишком большим,

так как качество результатов данного теста напрямую зависит от количества элементов, например:

Если алгоритм выдаст обратное ранжирование, то каждая последующая компонента теста будет все ближе подходить к 1, но это не будет означать, что алгоритм выдает хороший ответ.

## Kendall Test

**Описание** Заданные вектора нормализуются, а затем для них считается корреляция Кендалла.

**Оценка метрики** Тест оценивает совпадение последовательностей, при этом имеет схожие с Sequence Test недостатки, но при этом эти ошибки имеют куда менее заметный эффект. За это сглаживание приходится платить понятностью, метрика немного более сложная в интерпретации.

## 2.2 Запуск тестов

Чтобы запустить тесты, достаточно запустить метод CompleteTest() из модуля Test, то есть выполнить следующий код:

```
# Python
```

```
import PageRank.Test as t  
t.CompleteTest()
```

Результат тестов появится в файле CompleteTestResults.html

## Глава 3

# Результаты алгоритмов на тестах

### 3.1 Итеративные алгоритмы

Все итеративные алгоритмы выдают полностью правильные ответы на примерах (первый и третий тесты), при этом на втором тесте в ответах появляются расхождения из-за специфики работы тех или иных алгоритмов, но при этом лидирующие страницы определяются точно. Исключением является цепь Маркова, она имеет серьезные расхождения с Power Method и его улучшениями.

## 3.2 Алгоритмы Монте Карло

**Endpoint Random Algorithms** Алгоритмы данной группы показывают неплохие результаты на примерах, на втором тесте показатели хуже, но топ страниц определен достаточно точно, при этом они отработали примерно в 6 раз быстрее Extrapolating Adaptive Power Method.

**Complete Path Algorithms** Алгоритмы показывают не очень хорошие значения части тестов, что частично связано с малым заданным числом проходов (алгоритмы в тестах настроены на небольшое число страниц). Имеют преимущества предыдущей группы алгоритмов: скорость, высокое качество определения лидирующих страниц, но имеют лучшие значения тестов. В этой группе выделяется Stopping Complete Path Monte Carlo Method, так как все страницы без ссылок на них имеют одинаковое значение ранга.