Министерство образования и науки Российской Федерации

Федеральное государственное образовательное бюджетное учреждение

Тверской государственный технический университет

Кафедра “Программное обеспечение”

Курсовая работа

по теме: «Алгоритмы ранжирования. PageRank»

Выполнил студент

Группы «Б.ПИН.РИС.22.06»

Кириллов Владислав Константинович

Проверил:

Мальков Александ Анатольевич

Тверь. 2024

**Содержание**

Введение 3

Аналитическая часть 4

Что такое PageRank 4

Где используется 7

Проблемы алгоритма 8

Модификации алгоритма 9

Среда разработки Visual Studio 11

Проектирование 13

Логика приложения 13

Требования к приложению 14

Код алгоритма и этапы его работы 15

Тестирование 18

Документация 19

Заключение 21

Список литературы 22

Введение

Алгоритмы ранжирования играют важную роль в области информационного поиска, поскольку они определяют порядок, в котором результаты поиска должны быть представлены пользователю. Один из наиболее известных алгоритмов ранжирования - PageRank, разработанный Google.

PageRank является одним из многих факторов, которые Google использует для определения релевантности и качества веб-страницы. Он помогает поисковой системе предоставлять пользователям наиболее релевантные и качественные результаты поиска.

Цели работы:

1. Изучить алгоритм PageRank и его основные принципы работы.

2. Разработать программу для реализации алгоритма PageRank.

3. Провести тестирование разработанной программы.

4. Сделать выводы по результатам работы.

Задачи:

1. Изучить литературу и ресурсы в Интернете для понимания алгоритма PageRank и его основных принципов работы.

2. Определить требования и параметры для программы реализации алгоритма PageRank.

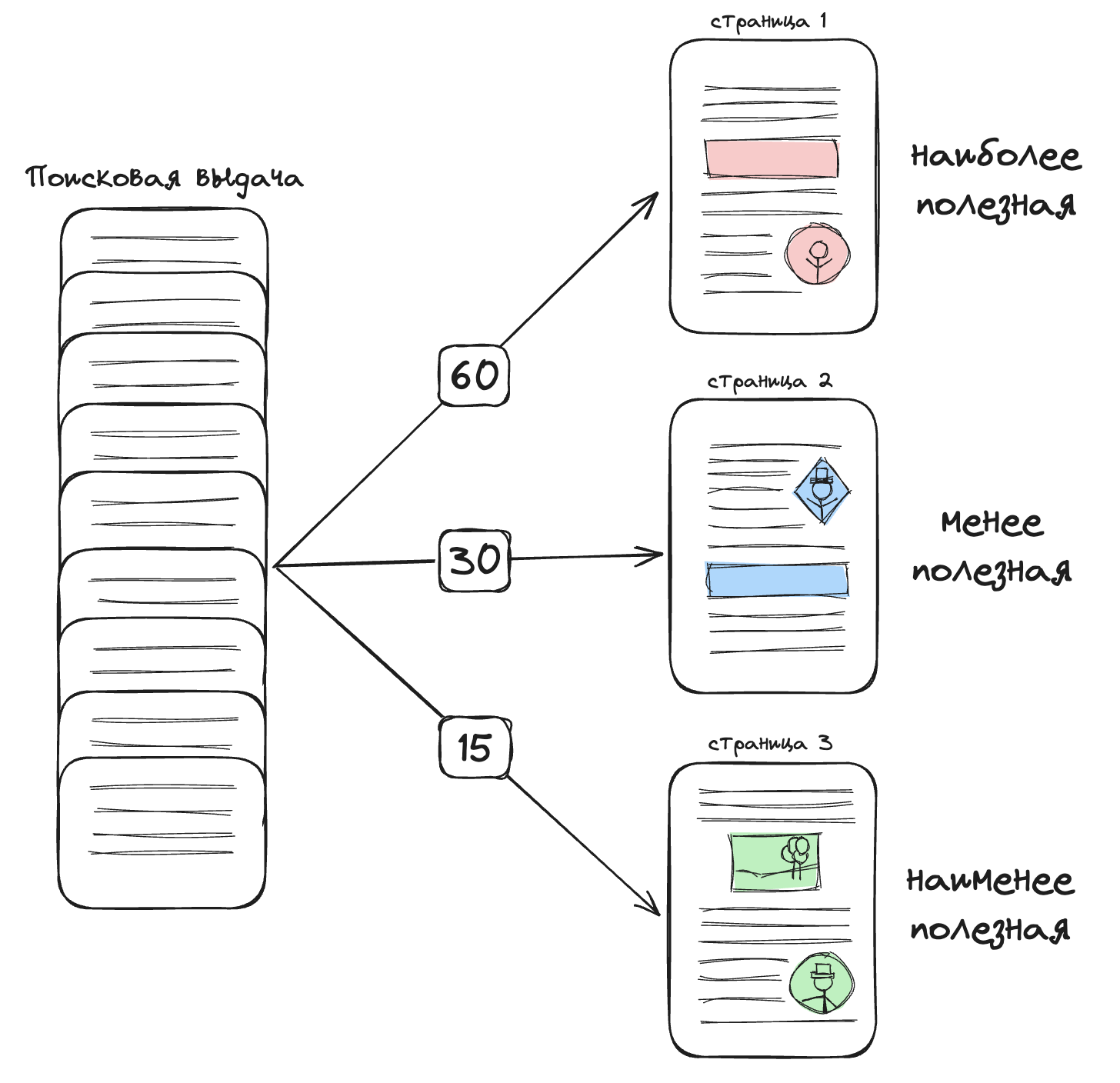
3. Написать код алгоритма PageRank.

4. Провести тестирование программы.

Аналитическая часть

**Что такое PageRank**

PageRank - это алгоритм, разработанный Ларри Пейджем и Сергеем Брином в 1998 году, который используется для ранжирования веб-страниц в поисковых системах. Он оценивает количество и качество ссылок, ведущих на страницу, и определяет ее авторитетность. Это был прорыв в области поиска, так как впервые поисковая система начала оценивать авторитетность страниц, что позволило улучшить качество результатов поиска.



Алгоритм строился с учетом *“случайного интернет пользователя”*, который переходит по страницам, нажимая по ссылкам. PageRank страницы — это вероятность, с которой этот интернет-пользователь перейдет на нее. Оценка рассчитывалась в пределах от 0 до 10: чем она выше, тем выше авторитет страницы.

PageRank — это попытка оценивать полезность веб-страницы в соответсвии с субъективным поведением пользователя. Чем чаще на страницу переходят, тем более она полезна для людей.

**Формула алгоритма**

Также PageRank учитывает авторитет ссылающихся источников. Таким образом, чем больше вес PageRank у одной страницы, тем больше вес она передаст странице, на которую она ссылается.

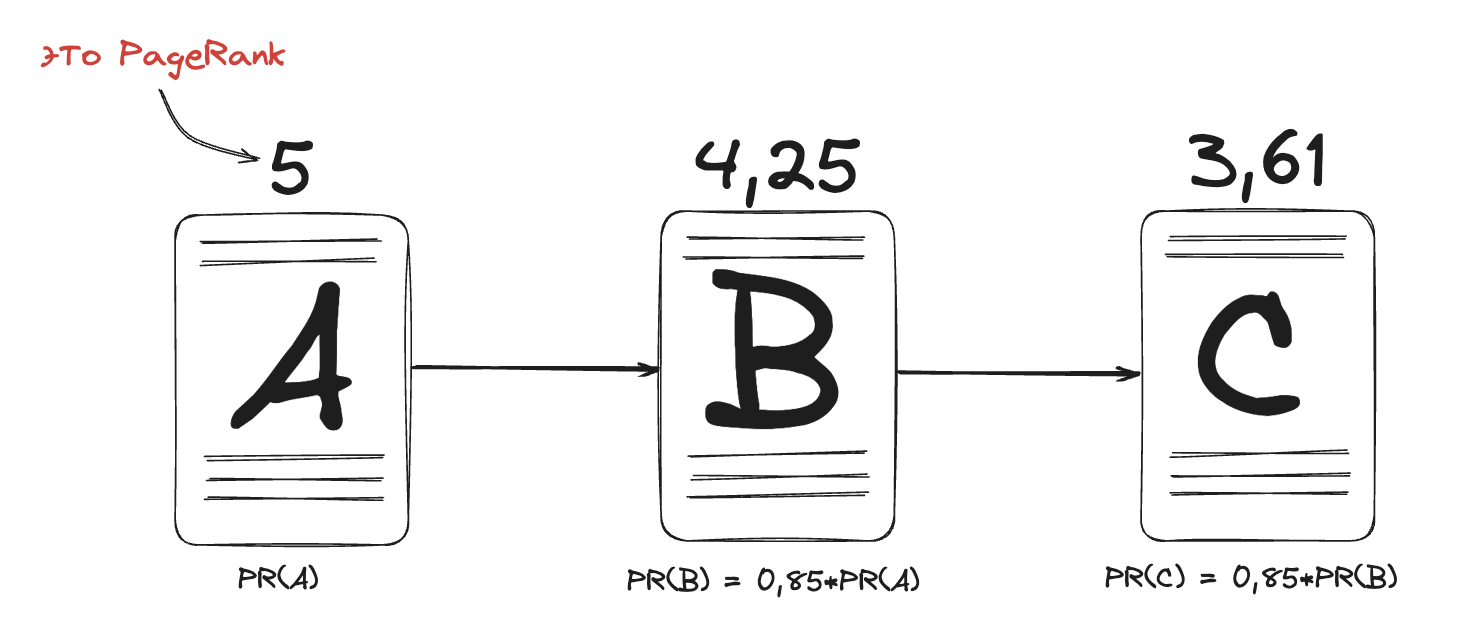
Алгоритм PageRank основан на математической формуле, которая описывает процесс распределения весов между веб-страницами. Формула выглядит следующим образом:

В данной формуле:

* А - исследуемая страница
* d - коэффициент затухания, то есть вероятность того, что пользователь покинет веб-страницу (обычно это 0.85).
* С - количество ссылок на исследуемой веб-странице
* Т1 … Тn - страницы, которые ссылаются на исследуемую страницу.

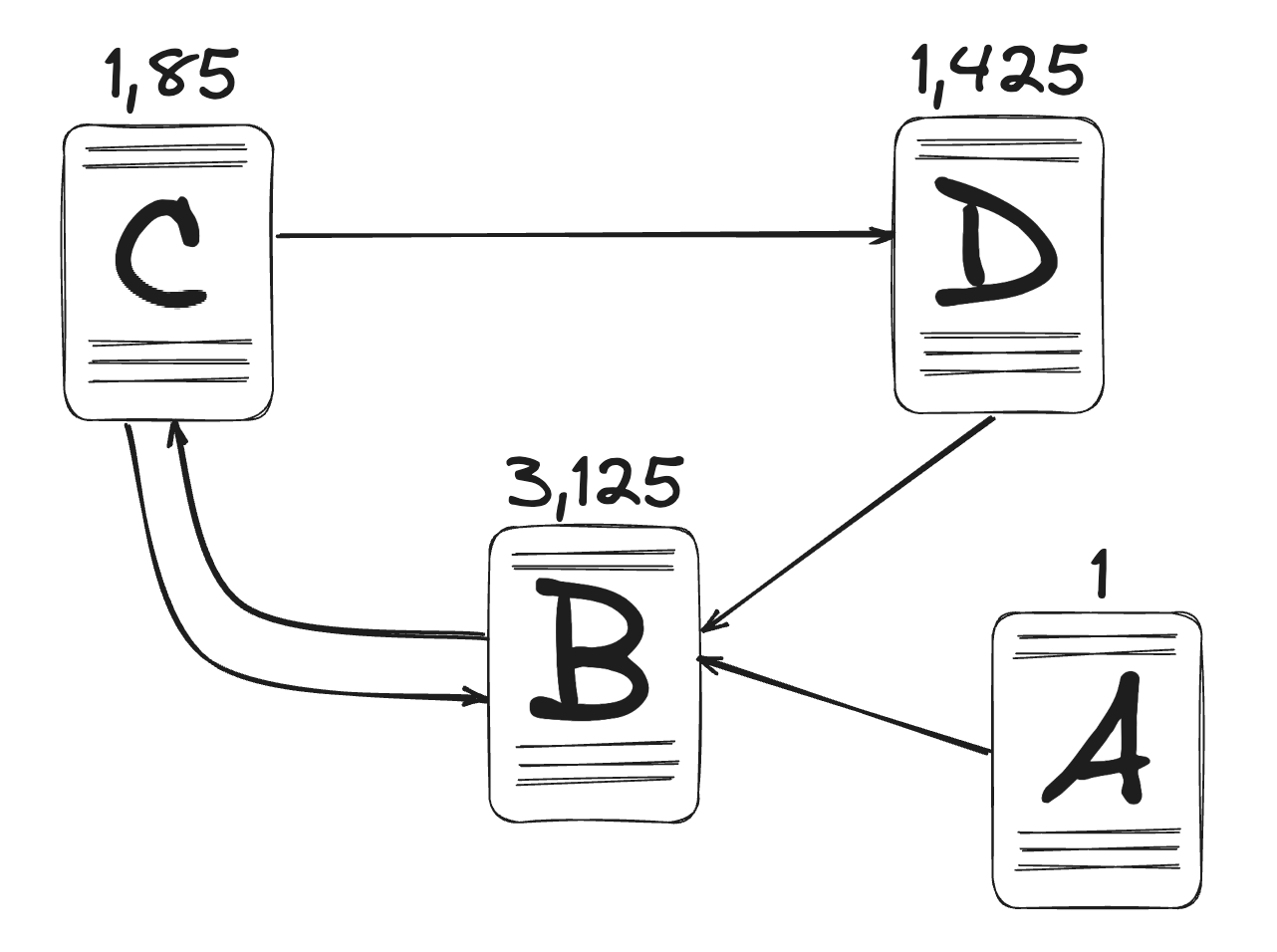
Согласно этой формуле, страницы распределяют свою оценку PageRank другим страницам, на которые они ссылаются.

**Пример 1:**



Например, источник A с оценкой 5 цитирует B. Не беря во внимания другие ссылки, которые может иметь в себе B, эта страница получает 85% от оценки A, то есть 4.25. Если при этом B цитирует С, оценка С будет составлять 85% от 2.125, и так далее.

**Пример 2:**



**Оценка сложности алгоритма**

Сложность алгоритма PageRank зависит от двух факторов: количества страниц в графе (n) и количества итераций, необходимых для сходимости (k). На каждой итерации алгоритм обновляет ранг каждой страницы, используя формулу, которая учитывает ранги соседних страниц и количество исходящих ссылок с них. Это занимает

времени, так как нужно пройти по всем элементам матрицы смежности графа. Таким образом, общая сложность алгоритма PageRank составляет

Количество итераций k зависит от выбранного порога сходимости и демпинг-фактора. Чем меньше порог, тем больше точность, но и больше итераций.

Демпинг-фактор - это вероятность того, что пользователь продолжит переходить по ссылкам, а не покинет страницу. Обычно он равен 0.85. Можно показать, что для достижения сходимости с порогом.

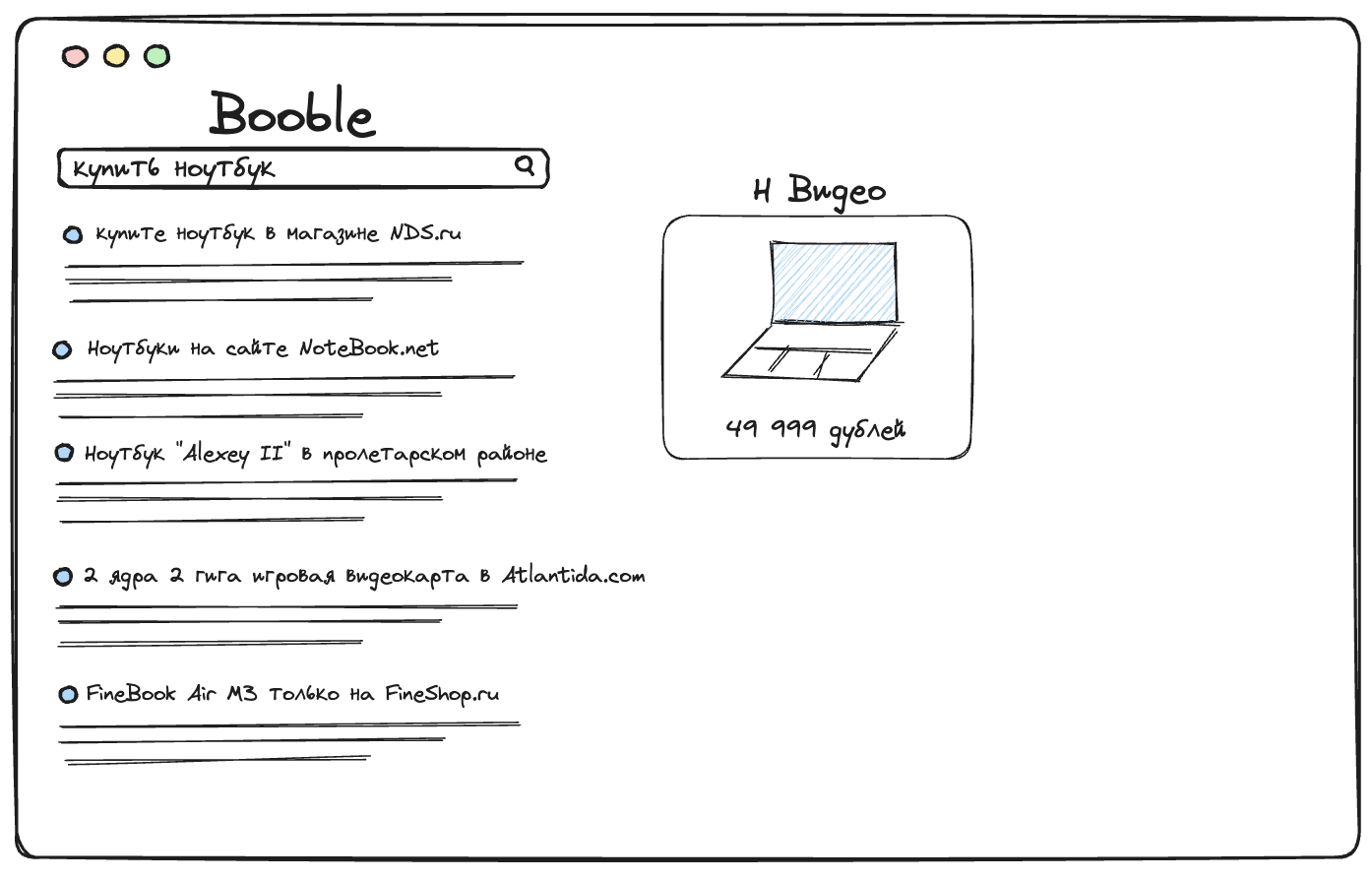
необходимо выполнить

Итераций.

Таким образом, сложность алгоритма PageRank можно также выразить как

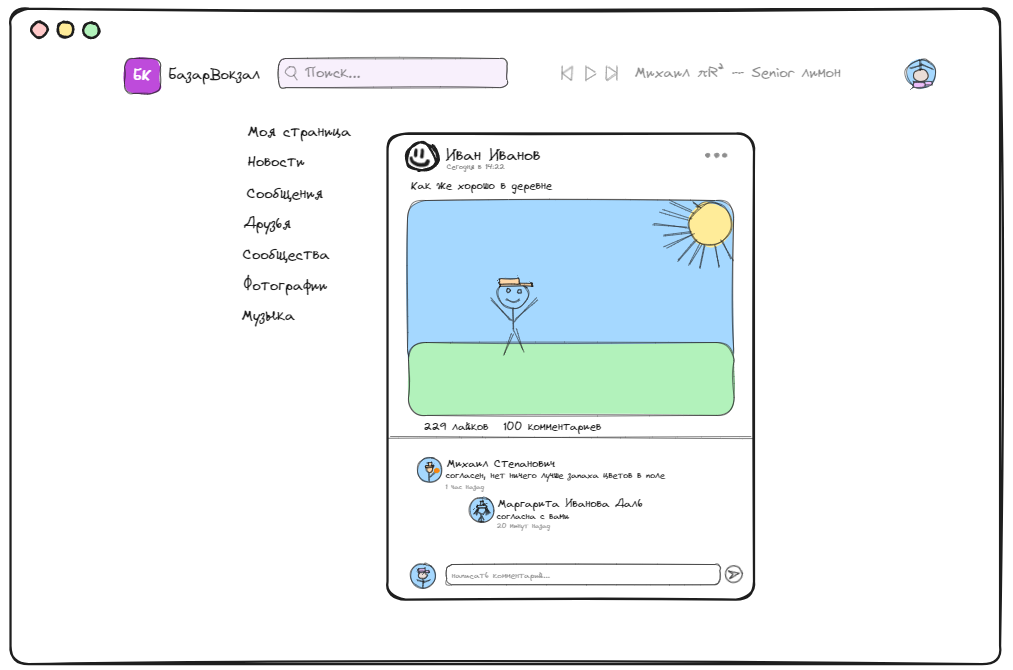
**Где используется**

**Поисковые системы**



В поисковых системах, таких как Google, PageRank используется для определения релевантности веб-страницы по запросу пользователя. Когда пользователь вводит запрос, поисковая система использует алгоритм PageRank для определения, какие веб-страницы наиболее важны и релевантны этому запросу.

**Социальные сети**



В социальных сетях PageRank может быть использован для определения важности и популярности постов, пользователей и групп. Он учитывает количество лайков, комментариев и репостов, которые получает пост, пользователь или группа, а также авторитетность пользователей, которые взаимодействуют с ними.

**Проблемы алгоритма**

Алгоритм PageRank имеет несколько проблем, которые могут влиять на его эффективность и точность. Некоторые из них включают:

1. Страницы без выходных ссылок (dead-end). Это может вызвать ситуацию, когда важность "утекает" с таких страниц.
2. Спам-ссылки: Алгоритм PageRank может быть подвержен манипуляциям со стороны веб-мастеров, которые создают большое количество спам-ссылок на свои веб-страницы с целью повышения их рейтинга. Это может привести к искажению результатов поиска.
3. Скрытые ссылки: Некоторые веб-мастера могут использовать скрытые ссылки, которые не видны пользователям, но которые все равно учитываются алгоритмом PageRank. Это также может привести к искажению результатов поиска.
4. Круговая ссылка: Если веб-страница ссылается только на другие страницы, которые ссылаются обратно на нее, это может привести к циклической ссылке, что может исказить результаты поиска.
5. Различные типы ссылок: Алгоритм PageRank учитывает только текстовые ссылки, но не учитывает другие типы ссылок, такие как изображения или кнопки. Это может привести к тому, что некоторые веб-страницы, которые имеют качественный контент, но не имеют текстовых ссылок, могут быть недооценены.
6. Недостаток обновлений: Алгоритм PageRank обновляется не так часто, как другие алгоритмы ранжирования, что может привести к тому, что некоторые веб-страницы, которые были изменены или улучшены, могут не получить соответствующего повышения рейтинга.

**Модификации алгоритма**

Алгоритм PageRank, созданный Сергеем Брином и Ларри Пейджем в 1998 году, претерпел ряд модификаций в процессе своего существования. Приведем несколько примеров таких модификаций:

1. Improved PageRank Algorithm - модификация, цель которой повышение точности алгоритма путем учета количества ссылок на страницу и их качества.
2. Weighted PageRank Algorithm - модификация, которая учитывает вес каждой ссылки, что позволяет более точно определить важность страницы.
3. Time-Dependent PageRank Algorithm - модификация, которая учитывает время активности страницы и ее популярность в определенный момент времени.
4. Social PageRank Algorithm - модификация, в рамках которой учитываются социальные связи между пользователями, что позволяет определить важность страницы на основе этих связей.
5. User Behavior-Based PageRank Algorithm - модификация, которая учитывает поведение пользователей на странице, такие как время пребывания на странице и количество кликов на ссылки.

Эти модификации алгоритма PageRank были предложены и исследованы в научном сообществе. Каждая из них обладает своими преимуществами и недостатками, и выбор конкретной модификации определяется в соответствии с конкретными требованиями и целями исследования.

Среда разработки Visual Studio

Visual Studio - это интегрированная среда разработки (IDE), созданная Microsoft. Она предлагает множество возможностей и инструментов для разработки программного обеспечения на различных языках программирования, включая C++, C#, Visual Basic, F#, JavaScript и Python.

Основные преимущества Visual Studio включают:

1. Интеграция с другими продуктами Microsoft: Visual Studio тесно взаимодействует с другими продуктами Microsoft, например, .NET Framework, Windows Presentation Foundation (WPF), Windows Communication Foundation (WCF) и Windows Workflow Foundation (WF).
2. Поддержка различных языков программирования: Visual Studio поддерживает множество языков программирования, делая ее универсальной средой разработки.
3. Отладка и профилирование: Visual Studio предлагает мощные инструменты для отладки и профилирования кода, помогающие разработчикам быстро находить и исправлять ошибки.
4. Интеграция с системой контроля версий: Visual Studio интегрирована с системой контроля версий Git, что позволяет разработчикам легко отслеживать изменения в коде.
5. Интеграция с облачными сервисами: Visual Studio взаимодействует с облачными сервисами Microsoft, такими как Azure, предоставляя разработчикам возможность легко развертывать и управлять своими приложениями в облаке.
6. Поддержка различных платформ: Visual Studio поддерживает разработку для различных платформ, включая Windows, Linux и macOS.
7. Удобный интерфейс: Visual Studio обладает интуитивно понятным и удобным интерфейсом, что позволяет разработчикам быстро освоиться и начать работу.

В общем, Visual Studio - это одна из самых мощных и универсальных IDE на рынке, которая предоставляет разработчикам множество возможностей и инструментов для создания высококачественного программного обеспечения.

Проектирование

**Логика приложения**

Описанная ниже логика приложения представляет собой последовательность действий, производимых в рамках реализации алгоритма PageRank:

1. Формирование входных данных: приложение обрабатывает веб-граф, представленный в форме списка веб-страниц и их взаимосвязей.
2. Поиск и хранение веб-страниц: приложение выполняет поиск и сохранение веб-страниц, содержащихся во входном графе. В этом контексте используется поисковой робот, выполняющий переход по всем страницам и извлекающий информацию о связях между ними.
3. Конструирование матрицы смежности: на основе полученной информации строится матрица смежности, в которой строки и столбцы представляют собой веб-страницы, а элементы матрицы указывают на наличие или отсутствие ссылок между страницами.
4. Реализация алгоритма PageRank: для построенной матрицы смежности выполняется процедура расчета ранжирования с использованием алгоритма PageRank. Данный алгоритм основан на принципе, согласно которому важность страницы определяется ее связями с другими веб-страницами, и количество ссылок, ведущих на данную страницу, коррелирует с ее рейтингом.
5. Вывод результатов: после применения алгоритма PageRank приложение представляет результаты ранжирования веб-страниц в порядке их значимости, что позволяет определить степень влияния конкретных веб-страниц в рамках данного веб-графа.

Итак, веб-граф подвергается обработке и анализу с использованием алгоритма ранжирования PageRank, что позволяет определить значимость и влияние конкретных веб-страниц в рамках графа.

**Требования к приложению**

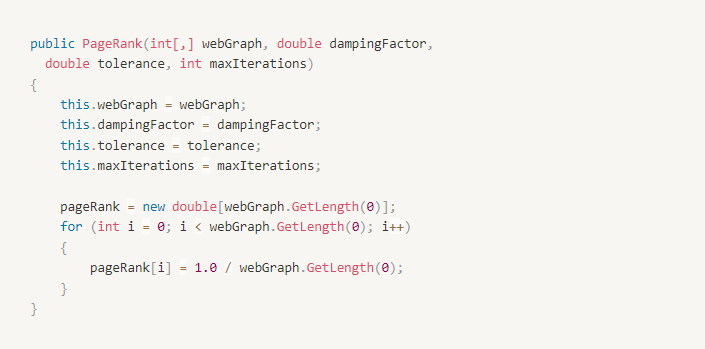
1. Приложение должно быть способно принимать на вход веб-графы, представленные в форме списка веб-страниц и их взаимосвязей.
2. Приложение должно быть способно обрабатывать и сохранять информацию о веб-страницах, содержащихся в входном графе.
3. Приложение должно создавать матрицу смежности на основе информации о веб-страницах и их взаимосвязях, где строки и столбцы представляют собой веб-страницы, а элементы матрицы указывают на наличие или отсутствие ссылок между страницами.
4. Приложение должно реализовывать алгоритм PageRank для построенной матрицы смежности и вычислять ранжирование веб-страниц.
5. Приложение должно представлять результаты ранжирования веб-страниц в порядке их значимости.
6. Приложение должно предусматривать проверку сходимости алгоритма по заданному порогу.
7. Приложение должно иметь графический интерфейс для ввода веб-графа и отображения результатов.
8. Приложение должно быть разработано на языке C# с использованием IDE Visual Studio.
9. Приложение должно быть документировано, включая комментарии к коду и пользовательскую документацию.

Код алгоритма и этапы его работы

Алгоритму ставится задача вычислить ранг веб-страниц, чтобы определить наиболее полезные и релевантные сайты, которые могут дать пользователю подходящий ответ на конкретный запрос. В данном случае запрос звучит так: "купить ноутбук". Цель - предложить пользователю те веб-страницы, которые будут наиболее полезными и информативными по данному запросу.

Алгоритм PageRank можно разделить на 5 этапов.

**Инициализация**

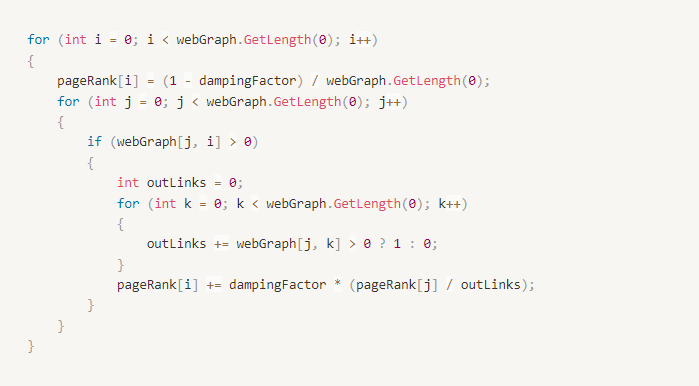


Для инициализации мы используем конструктор, которые принимает в себя 4 параметра:

1. webGraph — граф веб-страниц.
2. dampingFactor — определяет, какая часть веса передается на следующие страницы.
3. tolerance — определяет точность результата к окончательному значению.
4. maxIterations — максимальное количество итераций.

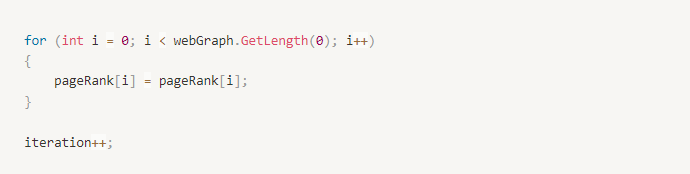
В конструкторе создается новый массив pageRank, который будет хранить значения PageRank для каждой страницы. Затем каждой странице присваивается начальный ранг, который обычно равен результату деления единицы на общее количество страниц в веб-графе.

**Распределение ранга**



На каждой итерации ранг каждой страницы распределяется между всеми страницами, на которые она ссылается. Если страница не имеет исходящих ссылок, её ранг равномерно распределяется между всеми страницами в графе.

**Обновление ранга**



После распределения рангов они обновляются для каждой страницы. Это включает в себя применение коэффициента затухания *(damping factor)*, чтобы учесть вероятность случайного перехода пользователя на другую страницу.

**Проверка сходимости**



После каждой итерации проверяется, изменились ли ранги страниц меньше, чем на заданный порог сходимости. Если изменения меньше порога, алгоритм завершает работу.

Тестирование

Attachment.png

Документация

**Характеристики используемой техники**

1. Тип: Персональный компьютер.
2. Версия операционной системы: Windows 11.
3. Тип процессора и его тактовая частота: Intel Core i5-10400F (4.30 ГГц)
4. Объем оперативной памяти: 16GB
5. Название и версия программного обеспечения: Visual Studio

**Краткое руководство пользователя**

1. Запустите приложение.
2. Ввод матрицы осуществляется в соответствующее текстовое поле. В случае необходимости использования случайной генерации, следует заполнить поле с размером матрицы.
3. Далее указывается коэффициент затухания, который отвечает за скорость затухания Page Rank.
4. Задаются параметры точности и максимального числа итераций. Точность ограничивает меру допустимого отклонения результатов вычислений Page Rank от истинного значения, максимальное число итераций определяет предельное количество итераций, которое приложение будет выполнять.
5. Процесс вычисления Page Rank инициируется нажатием кнопки "Получить Page Rank".
6. После активации кнопки приложение начинает процесс вычисления Page Rank для каждой страницы. Результаты представляются в текстовом поле в форме списка, где каждая строка соответствует странице, а столбец - ее значению Page Rank.
7. По завершении процесса вычисления можно получить итоговый результат.

Отмечу, что точность и максимальное число итераций оказывают существенное влияние на время вычисления Page Rank. Увеличение точности и числа итераций приводит к повышению точности результатов, однако увеличивает время вычисления.

Заключение

В заключении можно сказать, что алгоритм PageRank является важным инструментом для определения релевантности и качества веб-страниц в области информационного поиска. Разработка программы для реализации этого алгоритма позволяет проводить более точное ранжирование результатов поиска и предоставлять пользователям наиболее релевантные и качественные результаты. Тестирование программы подтвердило ее работоспособность и эффективность. Таким образом, использование алгоритма PageRank в разработке программного обеспечения для информационного поиска является перспективным направлением, которое может значительно улучшить качество поиска и удовлетворенность пользователей.

Список литературы

1. Wikipedia, "PageRank", <https://ru.wikipedia.org/wiki/PageRank>.
2. Klepikov, Konstantin, "PageRank: Алгоритм и его применение", <https://konstantinklepikov.github.io/2021/02/19/page-rank.html>.
3. Seranking, "PageRank: что это такое и как он работает", <https://seranking.com/ru/blog/pagerank/>.
4. Wikipedia, "PageRank", <https://en.wikipedia.org/wiki/PageRank>.
5. Liderpoiska, "Google PageRank: что это такое и как он работает", <https://liderpoiska.ru/blog/google-pagerank/>.