МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РФ

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ   
ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ   
«ВЯТСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»

Институт математики и информационных систем

ФАКУЛЬТЕТ КОМПЬЮТЕРНЫХ И ФИЗИКО-МАТЕМАТИЧЕСКИХ НАУК

КАФЕДРА ПРИКЛАДНОЙ МАТЕМАТИКИ И ИНФОРМАТИКИ

Допущен к защите

Заведующий кафедрой ПМИ

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Е.В. Разова

**Разработка системы ссылочного ранжирования**

Курсовой проект по дисциплине  
«Теория алгоритмов»

Выполнил студент группы ФИб-3301-51-00     / Д. О. Ощепков/

Руководитель к.п.н.,

старший преподователь кафедры ПМИ     / А. Г. Татаринова/

Работа защищена с оценкой     \_\_\_\_.\_\_\_\_.2024 г.

Члены комиссии:     / /

    / /

Киров 2023

Содержание

[Введение 3](#_Toc162293825)

[1 Разбор алгоритмов ссылочного ранижирования 5](#_Toc162293826)

[1.1 Постановка задачи 5](#_Toc162293827)

[1.2 Метрики оценки качества 5](#_Toc162293828)

[1.2.1 Precision и recall 5](#_Toc162293829)

[1.2.2 метрика и ее частный случай метрика 6](#_Toc162293830)

[1.2.3 метрика и ее частный случай метрика 7](#_Toc162293831)

[1.3 Вычислительная сложность рассматриваемых алгоритмов 7](#_Toc162293832)

[1.4 Pagerank 7](#_Toc162293833)

[1.4.1 Формулировка 7](#_Toc162293834)

[1.4.2 Достоинства и недостатки 7](#_Toc162293835)

[1.4.3 Алгоритм 7](#_Toc162293836)

[1.5 Выводы по разделу 7](#_Toc162293837)

[2 Применение алгоритмов на практике 8](#_Toc162293838)

[2.1 Какие-то данные 8](#_Toc162293839)

[2.2 Выводы по разделу 8](#_Toc162293840)

[Заключение 9](#_Toc162293841)

[Библиографический список 10](#_Toc162293842)

[Приложения 11](#_Toc162293843)

[Приложение А. Листинг программы 11](#_Toc162293844)

# Введение

Основная идея алгоритмов ссылочного ранжирования заключается в том, чтобы определить степень важности каждого элемента на основе структуры графа, образованного самими элементами и связями между ними.

Эти алгоритмы часто применяются в поисковых системах для ранжирования результатов поиска (<https://papers.cumincad.org/data/works/att/2873.content.pdf> ), в социальных сетях для определения влиятельных пользователей или контента (<https://link.springer.com/article/10.1007/s13278-014-0172-x> ), в анализе сетей для выявления ключевых узлов (<https://www.mdpi.com/1996-1073/15/3/797> )

Основные алгоритмы ссылочного ранжирования:

1. Pagerank и его вариации, такие как personalized pagerank, weighted pagerank, personalized weighted pagerank. Они используют итеративный процесс для расчета весов страницы, учитывая количество и качество входящих ссылок. (<https://papers.cumincad.org/data/works/att/2873.content.pdf> )
2. Hits и его вариации. Используется для определения важности веб-страниц на основе их характеристик как "хабов" и "авторитетов". Хабы (Hubs) - это страницы, которые содержат много ссылок на другие страницы. Чем больше ссылок у страницы, тем она более "хабовая". Авторитеты (Authorities) - это страницы, на которые много ссылается. Такие страницы считаются важными и качественными и обладают высоким "авторитетом". (<https://dl.acm.org/doi/pdf/10.1145/324133.324140> )

Проблема исследования заключается в отсутствии четких рекомендаций или наставлений для выбора конкретного алгоритма ссылочного ранжирования классификации в конкретной задаче. Каждый из методов обладает своими сильными и слабыми сторонами, а также применим к определенным условиям и задачам.

Для решения данной проблемы необходимо провести сравнительный анализ различных алгоритмов ссылочного ранжирования в различных сценариях и задачах. Это позволит определить, какие методы лучше справляются с определенными типами данных, объемом данных, структурой сети и характеристиками узлов. Кроме того, такой анализ может учитывать требования к точности ранжирования, скорости работы алгоритма

Целью курсового проекта является формирование рекомендации по выбору модели.

Для достижения цели необходимо решить следующие задачи.

1. Обзор литературы. Изучить существующие научные статьи, книги и исследования, посвященные метрическим методам классификации. Выявить основные методы и их характеристики, а также рассмотреть примеры исследований, где эти методы применялись в различных задачах.
2. Описание и анализ методов ранжирования
3. Подготовка набора данных. Собрать или использовать доступные наборы данных, соответствующие различным предметным областям. Обеспечить разнообразие данных в терминах признаков, классов, объема выборки и сложности задачи.
4. Реализация и применение каждого из методов для решения задач машинного обучения на различных наборах данных.
5. Анализ результатов и сравнение эффективности каждого метода на различных наборах данных. Определить подходящие метрики оценки качества классификации и провести серию экспериментов, сравнивая различные метрические методы классификации на подготовленных наборах данных. Сравнивать производительность методов в терминах точности, полноты, F-меры.
6. Сделать выводы о применимости каждого метода в различных сценариях и задачах ранжирования.

Курсовой проект включает два раздела. В первом разделе проведен обзор литературы и алгоритмов по предметной области. Во втором разделе был проведен анализ алгоритмов ранжирования.

# Разбор алгоритмов ссылочного ранижирования

## Постановка задачи

## Метрики оценки качества

Формализуем человеческую оценку в функциональный вид

(<https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=0bee052af002eb197277cd222d62154c7de4ac8a> , 578)

Где , а – множество вершин графа страниц

Хорошая страница или плохая можно определить по разному на практике. Этой оценкой может быть, например, то, что значение, даваемое экспертом, выше чем среднее значение по всем оценкам страниц.

### Precision и recall

Precision (точность) измеряет, насколько точно модель классифицирует положительные примеры. Она вычисляется как отношение числа правильно классифицированных положительных примеров к общему числу примеров, которые модель отнесла к положительному классу. Для класса :

Recall (полнота) измеряет, насколько хорошо модель обнаруживает все положительные примеры. Она вычисляется как отношение числа правильно классифицированных положительных примеров к общему числу положительных примеров в данных. Для класса :

Таким образом, precision полезна в задачах, где важно минимизировать ложно положительные результаты, а recall полезна в задачах, где важно минимизировать ложно отрицательные результаты.

Выбор между этими двумя метриками зависит от исходной задачи.

Одним из примеров, когда стоит предпочесть одну метрику другой, может быть задача классификации медицинских изображений для обнаружения редких заболеваний. В таком случае, полнота (recall) может иметь большее значение, поскольку нежелательно пропустить даже один положительный случай.

Напротив, если классификация используется для фильтрации спама в электронной почте, точность (precision) может иметь большее значение. В этом случае, важно минимизировать количество ложно положительных предсказаний, чтобы не пропустить важные сообщения, помечая их как спам.

### метрика и ее частный случай метрика

Является объединением Precision и Recall, показывает приоритет одной метрики над другой. При метрики учитываются в равной степени. [9, p. 33, 36]

-мера полезна если данные несбалансированны, так как она учитывает и точность, и полноту, что позволяет получить общую оценку производительности модели, не учитывая смещение в сторону преобладающего класса.

### метрика и ее частный случай метрика

Accuracy вычисляется как отношение числа правильно классифицированных примеров к общему числу примеров:

[10, стр. 71]

Accuracy может быть полезна, когда классы в данных сбалансированы, то есть количество примеров каждого класса примерно одинаково. В таких случаях она дает общую оценку производительности модели.

## Вычислительная сложность рассматриваемых алгоритмов

## Pagerank

### Формулировка

### Достоинства и недостатки

### Алгоритм

## Выводы по разделу

# Применение алгоритмов на практике

## Какие-то данные[[1]](#footnote-1)

## Выводы по разделу

# Заключение

# Библиографический список

# Приложения

## Приложение А. Листинг программы

1. [↑](#footnote-ref-1)