НОВОСИБИРСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ФАКУЛЬТЕТ ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ

# Теоретические основы обработки информации

## Алгоритм Pagerank

Кондренко К.П, Неретин С.И., группа 21203

2 апреля 2024 г.

# Содержание

# 1 Введение

Ранжирование веб-страниц является одним из ключевых аспектов поисковых систем, так как позволяет предоставлять пользователям наиболее релевантные результаты в ответ на их запросы. Алгоритм PageRank является одним из наиболее известных и широко используемых методов ранжирования веб-страниц на основе их связности ссылками. Однако существуют и другие подходы к решению этой задачи, которые могут быть эффективны в различных контекстах. В данном обзоре рассматриваются другие методы ранжирования веб-страниц, их особенности, преимущества и недостатки.

# 2 Описание алгоритма PageRank

Алгоритм PageRank является одним из основных алгоритмов ранжирования веб-страниц, который используется в поисковых системах для определения их важности и релевантности для конкретного запроса пользователя. Он был разработан Ларри Пейджем и Сергеем Брином при создании поисковой системы Google.

В основе алгоритма лежит представление веба в виде графа, где узлами являются веб-страницы, а рёбра представляют собой гиперссылки между этими страницами. PageRank рассматривает каждую страницу как узел в этом графе и присваивает ей числовую оценку, которая отражает ее важность.

Принцип работы алгоритма можно описать следующим образом:

1. **Инициализация рангов**: В начале каждая страница инициализируется с одинаковым начальным рангом. Это может быть равномерное распределение или другой метод.

2. **Распространение ранга**: Затем выполняется итерационный процесс, в ходе которого ранги страниц обновляются на основе их собственного ранга и рангов страниц, которые на них ссылаются. Более важные страницы (те, которые имеют больше входящих ссылок) передают больше своего ранга страницам, на которые они ссылкаются.

3. **Учет демпинг фактора**: Для предотвращения переполнения и улучшения качества ранжирования вводится демпинг фактор (обычно около 0.85). Он определяет вероятность того, что пользователь перейдет на другую страницу, вместо того чтобы продолжать переходить по ссылкам.

4. **Конвергенция**: Процесс итераций продолжается до тех пор, пока ранги страниц не стабилизируются, то есть до тех пор, пока изменения рангов страниц между итерациями не станут незначительными.

Важно отметить, что PageRank не является единственным фактором ранжирования, используемым в поисковых системах, но он является одним из

ключевых и может влиять на позицию веб-страниц в результатах поиска. Алгоритм был дополнен и улучшен с течением времени, чтобы более точно отражать реальные потребности пользователей и бороться с различными видами спама и манипуляций.

# 3 Специфика задачи ранжирования

Ранжирование веб-страниц является критически важной задачей для поисковых систем, однако аналогичные принципы ранжирования могут быть применены и в других областях, например, в обработке текстов. Один из примеров такого применения - алгоритм Textrank, который используется для автоматического создания кратких содержаний текстов.

# 4 О Алгоритме Textrank и как он основан на Pagerank

Textrank основан на предположении, что важные предложения в тексте часто содержат ключевую информацию и, следовательно, могут рассматриваться как важные "страницы"в графе предложений. Принцип работы Textrank аналогичен алгоритму PageRank, где вместо веб-страниц рассматриваются предложения текста, а вместо ссылок - связи между предложениями. Таким образом, Textrank позволяет определить наиболее важные предложения в тексте на основе их взаимосвязей.

# 5 Анализ существующих подходов

## 5.1 HITS

Алгоритм HITS (Hyperlink-Induced Topic Search) также используется для ранжирования веб-страниц, но в отличие от PageRank, HITS оценивает не только важность страницы как источника информации (авторитетности), но и как источника ссылок (хабовости). Кратко описывая алгоритм HITS:

1. Инициализация всех страниц как хабовых и авторитетных.
2. Итеративное обновление хабовости и авторитетности страниц.
3. Рекурсивное обновление рангов до сходимости.
4. Получение конечных значений хабовости и авторитетности для каждой страницы.

Разница между алгоритмами заключается в том, что PageRank оценивает важность страницы на основе ее собственных характеристик и характеристик страниц, которые на нее ссылаются, в то время как HITS учитывает как важность страницы, так и ее активность как источника ссылок.

## 5.2 BM25

BM25 (Best Matching 25) представляет собой вероятностный метод ранжирования, используемый в информационном поиске для оценки релевантности документов по отношению к запросу пользователя. Этот алгоритм является усовершенствованным вариантом модели TF-IDF (Term Frequency-Inverse Document Frequency), который учитывает не только частоту терминов в документе, но и другие факторы, такие как длина документа и отдельные компоненты запроса.

Важным преимуществом BM25 является его способность эффективно обрабатывать как короткие, так и длинные текстовые документы и запросы. В отличие от алгоритмов ранжирования веб-страниц, таких как HITS и PageRank, которые оценивают важность веб-страниц на основе их авторитетности и структуры ссылок между ними, BM25 оценивает релевантность документов на основе совпадения между терминами запроса и содержанием документа, учитывая различные факторы для достижения наилучшего соответствия запросу.

## 5.3 Системы на основе машинного обучения

С развитием методов машинного обучения появились и методы ранжирования, основанные на обучении моделей на больших наборах данных. Например, вместо использования только структуры графа ссылок, данные о поведении пользователей и текстовое содержимое страниц могут быть использованы для создания модели ранжирования. Методы машинного обучения также могут учитывать контекст запроса пользователя и другие факторы, что делает их более адаптивными к изменяющимся требованиям пользователей.

# 6 Актуальность и специфика подходов

Каждый из перечисленных подходов имеет свои преимущества и недостатки, а также свою область применения. Например, алгоритм PageRank хорошо работает на больших графах сильно связанных страниц, но может быть менее эффективен на страницах с малым количеством ссылок. Алгоритм HITS может быть более подходящим для поиска конкретных тематических сообществ в веб-графе.

Системы на основе машинного обучения обычно требуют большого количества данных для обучения, но могут быть более гибкими и точными в ранжировании. Они также могут лучше адаптироваться к изменяющимся требованиям пользователей и окружению. Использование данных из социальных сетей может повысить релевантность ранжирования, учитывая социальные факторы и поведение пользователей.

# 7 Заключение

В заключение, ранжирование веб-страниц является сложной задачей, и существует множество подходов к ее решению. Каждый из этих подходов имеет свои особенности, преимущества и недостатки. Выбор подхода зависит от конкретных требований и условий задачи. Дальнейшие исследования в этой области могут привести к созданию более эффективных методов ранжирования и улучшению качества поисковых систем.

# 8 Литература


- Brin, S., Page, L. (1998). The Anatomy of a Large-Scale Hypertextual Web Search Engine.
- Kleinberg, J. M. (1999). Authoritative sources in a hyperlinked environment.
- Langville, A. N., & Meyer, C. D. (2006). Google's PageRank and beyond: The science of search engine rankings.
- Liu, B. (2011). Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data.
- Manning, C. D., Raghavan, P., & Schütze, H. (2008). Introduction to information retrieval.
- Richardson, M., & Domingos, P. (2002). The intelligent surfer: Probabilistic combination of link and content information in PageRank.
- Wu, S., & Wen, J. R. (2011). Learning to rank for information retrieval.
- https://habr.com/ru/articles/533096/