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 **3**

2 О алгоритме **3**

3 Специфика задачи ранжирования и TextRank **3**
3.1 Идея TextRank 3

4 Анализ существующих подходов **4**
4.1 HITS 4
4.2 BM25 4
4.3 Системы на основе машинного обучения 5

5 Актуальность и специфика подходов **5**

6 Заключение **5**

7 Литература **6**

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 О алгоритме

Алгоритм PageRank разработан Ларри Пейджем и Сергеем Брином в рамках исследования по созданию поисковой системы Google. Основная идея алгоритма заключается в том, чтобы оценить важность веб-страницы на основе количества ссылок, указывающих на нее, и важности самих страниц, ссылающихся на нее. Это делает его алгоритмом ранжирования, основанным на графе ссылок. Вкратце, алгоритм PageRank работает следующим образом:

3 Специфика задачи ранжирования и TextRank

Задача ранжирования веб-страниц является чрезвычайно важной для поисковых систем, но принципы ранжирования могут быть применены и в других областях, таких как обработка текстов. Например, в задаче автоматического создания краткого содержания для текстов TextRank использует принципы алгоритма PageRank для оценки важности предложений в тексте.

1. Инициализация всех страниц равной вероятностью.
2. Рекурсивное вычисление PageRank для каждой страницы на основе входящих ссылок и их весов.
3. Итеративное обновление PageRank до сходимости.
4. Получение конечных значений PageRank для каждой страницы.

3.1 Идея TextRank

TextRank основан на идее, что важные предложения в тексте часто содержат ключевую информацию и могут быть рассмотрены как важные "страницы"в графе предложений. Принцип работы TextRank аналогичен алгоритму PageRank:

4 Анализ существующих подходов

4.1 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.

4.2 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.

4.3 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.

5 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.

6 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.

7 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/