УДК 004.04

Эм Александра Андреевна,
студентка магистратуры
Международный Университет Информационных Технологий
Казахстан, г. Алматы
e-mail: allleksandraem@gmail.com

Научный руководитель: Сербин Василий Валерьевич
к.т.н., ассоциированный профессор
Международный Университет Информационных Технологий
Казахстан, г. Алматы

ПОИСК ТЕКСТОВЫХ ДОКУМЕНТОВ В СЛАБОСТРУКТРУРИРОВАННЫХ ИНФОРМАЦИОННЫХ МАССИВАХ

Аннотация: *В статье рассматривается проблема поиска текстовых документов в слабоструктурированных информационных массивах. Описаны методы, которые используются для поиска файлов определенных типов, таких как PDF или документы Word. Эти методы включают индексацию содержимого файлов, фильтрацию результатов поиска по типу файла, использование метаданных, специализированные алгоритмы и машинное обучение. Подробное изучение этих методов может помочь в создании более эффективных и точных поисковых систем.*

Ключевые слова: поисковые системы, поиск, индексация, документ, текст.

Em Alexandra Andreevna
master student
International Information Technology University,
Kazakhstan, Almaty

Scientific adviser: Serbin Vasily Valerievich
PhD, associated professor
International Information Technology University,
Kazakhstan, Almaty

SEARCH TEXT DOCUMENTS IN POORLY STRUCTURED INFORMATION ARRAYS

Abstract**:** *The article deals with the problem of searching text documents in weakly structured information arrays. The methods used to search for specific file types, such as*

PDF or Word documents, are described. These methods include indexing file contents, filtering search results by file type, using metadata, specialized algorithms, and machine learning. A detailed study of these methods can help in creating more efficient and accurate search engines.

Key words: search systems, search, indexing, document, text.

В современную цифровую эпоху мы ежедневно сталкиваемся с огромным количеством информации, которая генерируется и распространяется. Данные, с которыми мы сталкиваемся, часто слабоструктурированы и разбросаны по множеству источников, что затрудняет поиск конкретной информации. Эта проблема становится особенно сложной, когда нам нужно найти конкретный текстовый документ в слабоструктурированном массиве информации.

Поиск текстового документа в плохо структурированном информационном массиве может оказаться сложной задачей. Однако существуют некоторые методы и инструменты, которые могут помочь нам упростить этот процесс и повысить точность результатов поиска.

Чтобы повысить точность результатов поиска используются расширенные методы поиска, такие как булевы операторы, подстановочные знаки и поиск по близости. Булевы операторы позволяют комбинировать поисковые термины для уточнения результатов поиска. Подстановочные знаки позволяют искать вариации слова, например, различные написания или времена. Поиск по близости позволяет искать термины, которые находятся рядом друг с другом, что может быть особенно полезно при поиске фраз. Еще один инструмент, который может быть полезен при поиске текстового документа в слабоструктурированном информационном массиве - это поисковая система, специально разработанная для поиска текстовых документов. Такие поисковые системы предназначены для поиска конкретных типов содержимого, например, PDF-файлов или документов Word, и часто могут выполнять более сложный поиск, чем поисковые системы общего назначения.

Существует несколько методов, разработанными поисковыми системами для поиска определенных типов контента, таких как PDF или документы Word. Рассмотрим некоторые из них:

- Индексация содержимого файлов. Одним из основных методов, используемых поисковыми системами для поиска файлов определенных типов, является индексация содержимого файлов. При этом, поисковая система сканирует содержимое каждого файла и создает индекс, содержащий информацию о всех ключевых словах, фразах и других метаданных, которые могут помочь идентифицировать файл. Этот индекс затем используется для выполнения запросов поиска.

- Фильтрация результатов поиска по типу файла. Другой метод, который используется специализированными поисковыми системами для поиска файлов определенных типов, это фильтрация результатов поиска по типу файла. Поисковая система может быть настроена таким образом, чтобы искать только файлы в определенных форматах, таких как PDF или документы Word, и отфильтровывать все остальные типы файлов.

- Использование метаданных. Поисковые системы могут также использовать метаданные для идентификации файлов определенных типов. Например, PDF-файлы содержат специальные метаданные, такие как заголовки, авторы, даты создания и т.д. Поисковая система может использовать эту информацию для идентификации PDF-файлов и для облегчения поиска по этим файлам.

- Использование специализированных алгоритмов. Специализированные поисковые системы могут использовать специальные алгоритмы для поиска определенных типов контента.

- Использование машинного обучения. Некоторые поисковые системы могут использовать машинное обучение для облегчения поиска файлов определенных типов. Алгоритмы машинного обучения могут обучаться на больших

наборах данных и использовать эти данные для определения наиболее эффективных методов поиска файлов определенных типов.

Это лишь некоторые методы, которые могут использоваться специализированными поисковыми системами для поиска.

В заключение следует отметить, что поиск текстового документа в слабоструктурированном информационном массиве может оказаться сложной задачей. Однако, используя передовые методы поиска, специализированные поисковые системы и следуя лучшим практикам, можно повысить точность результатов поиска и найти нужную информацию более эффективно.