

ПОСТРОЕНИЕ БАЗЫ ДАННЫХ ПОИСКОВОЙ СИСТЕМЫ

Е.М. Кузнецова¹, Н.Ю. Юдина¹

¹ФГБОУ ВО «Воронежский государственный лесотехнический университет имени Г.Ф. Морозова»

Аннотация. В статье рассматривается вопрос создания базы данных поисковой системы. К основным аспектам математического описания базы данных поисковой системы относятся: описание коллекции документов, обратный индекс, модель запроса, оценка релевантности, ранжирование документов. База данных поисковой системы включает следующие таблицы: таблица «документы», таблица «слова», таблица «постинг-лист», таблица «связи», таблица «пользовательские запросы», таблица «результаты поиска». Кроме этого, в базу данных входят запросы, а также создаются индексы для увеличения скорости производительности.

Ключевые слова: база данных, индекс, модель, запросы, коллекция документов.

BUILDING A SEARCH ENGINE DATABASE

E.M. Kuznetsova¹, N.Yu. Yudina¹

¹Voronezh State University of Forestry and Technologies named after G.F. Morozov

Abstract. The article considers the issue of creating a search engine database. The main aspects of the mathematical description of the search engine database include: description of the document collection, reverse index, query model, relevance assessment, document ranking. The search engine database includes the following tables: the "documents" table, the "words" table, the "posting list" table, the "links" table, the "user queries" table, the "search results" table. In addition, the database includes queries, and indexes are created to increase the performance speed.

Keywords: database, index, model, queries, document collection.

База данных поисковой системы – это сложная структура, состоящая из множества компонентов, которые обеспечивают хранение, обработку и доступ к информации. Рассмотрим пример простой базы данных поисковой системы, включающей основные элементы.

К основным аспектам математического описания базы данных поисковой системы относятся: описание коллекции документов D , обратный индекс $I(w)$, модель запроса q , оценка релевантности $Relevance(d_i, q)$, ранжирование документов $Rank(d_i)$.

Множество просматриваемых документов можно описать (1) как

$$D = \{d_1, d_2, \dots, d_n\} \quad (1)$$

где n – количество документов в коллекции.

Каждый просматриваемый документ можно рассматривать как набор слов или текстовая строка (2).

$$d_i = \{w_1, w_2, \dots, w_{m_j}\} \quad (2)$$

здесь w_j – слова, которые содержатся в i -ом документе; m_j – количество слов в документе.

Взаимодействие этих компонентов позволяет эффективно индексировать, искать и ранжировать документы, обеспечивая высококачественные результаты поиска для пользователя.

Для быстрого поиска в структуре базы данных поисковой системы используется обратный индекс:

$$I(w) = \{(d_i, f_{wi}) | w \in d_i\} \quad (3)$$

где w – слово из словаря, d_i – просматриваемый документ, в котором встречается искомое слово, f_{wi} – частота появления слова.

Для каждого слова обратный индекс хранить набор документов, в которых это слова встречается.

Ниже приведено создание индексов на ключевые слова (рисунок 1).

```
CREATE INDEX idx_word ON Words(word);
CREATE INDEX idx_postings ON Postings(document_id, word_id);
CREATE INDEX idx_links_source ON Links(source document id);
CREATE INDEX idx_links_target ON Links(target document id);
```

Рисунок 1 – Создание индексов по ключевым полям

Степень соответствия найденного документа запросу определяется оценкой релевантности, которая определяется по формуле (4).

$$Relevance(d_i, q) = \sum_{w \in q} TF - IDF(w, d_i) \quad (4)$$

где $TF - IDF(w, d_i)$ – оценка релевантности с использованием модели TF-IDF; d_i – просмотренный документ.

После вычисления оценки релевантности всех просмотренных документов в коллекции выполняется упорядочивание по убыванию. Сортировка осуществляется в соответствии с формулой (5).

$$Rank(d_i) = sort(R(d_1), R(d_{12}), \dots, R(d_n)) \quad (5)$$

База данных поисковой системы включает следующие таблицы: таблица «документы», таблица «слова», таблица «постинг-лист», таблица «связи», таблица «пользовательские запросы», таблица «результаты поиска». Кроме этого, в базу данных входят запросы – поиск документов по слову и получение связанных запросов, а также создаются индексы для увеличения скорости производительности. Схема базы данных представлена на рисунке 2.

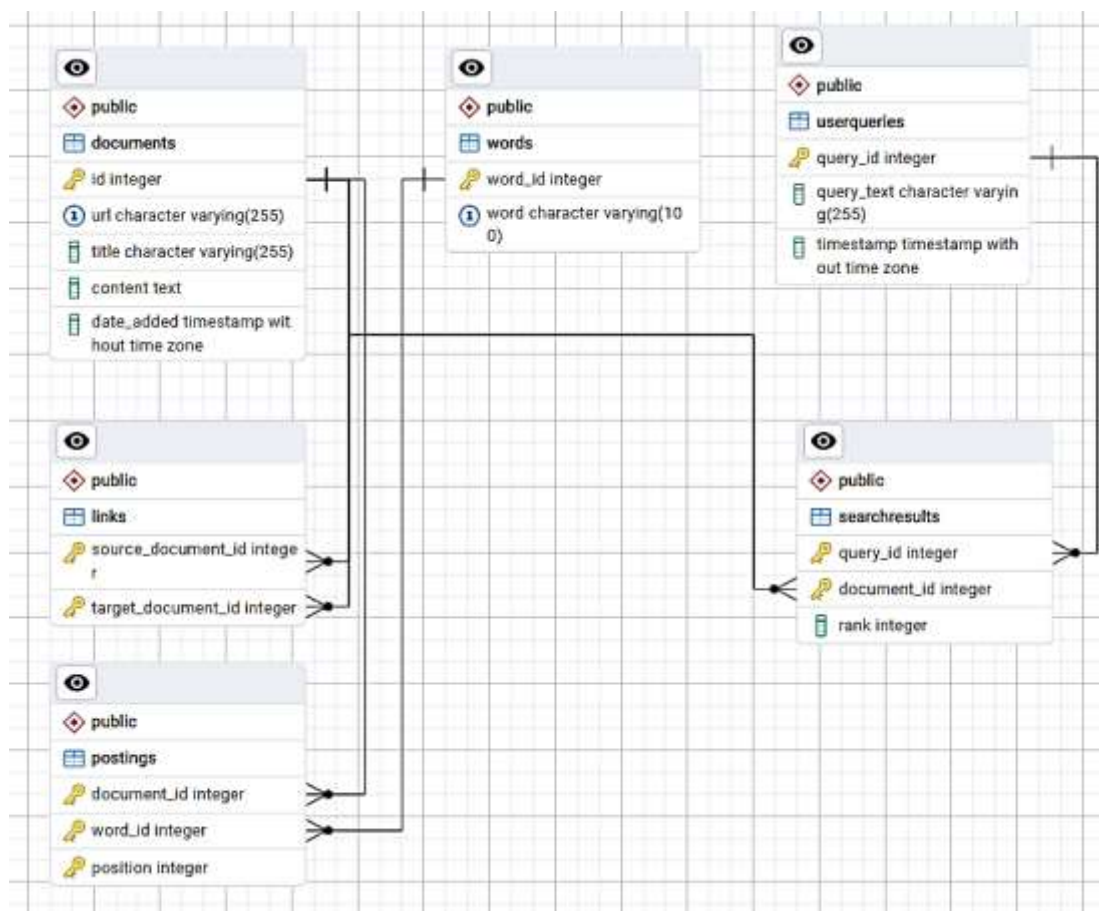


Рисунок 2 – Схема базы данных поисковой системы.

Примеры создания таблиц создаваемой базы данных. Таблица Documents содержит информацию о всех проиндексированных документах.

```
CREATE TABLE Documents (  
    id integer NOT NULL,  
    url character varying(255) COLLATE pg_catalog."default" NOT NULL,  
    title character varying(255) COLLATE pg_catalog."default",  
    content text COLLATE pg_catalog."default",  
    date_added timestamp without time zone DEFAULT CURRENT_TIMESTAMP,  
    CONSTRAINT documents_pkey PRIMARY KEY (id),  
    CONSTRAINT documents_url_key UNIQUE (url)  
);
```

Таблица Words хранит уникальные слова, которые встречались в документах.

```
CREATE TABLE Words (  
    word_id INT PRIMARY KEY,  
    word VARCHAR(100) NOT NULL UNIQUE  
);
```

Таблица Postings содержит связь между словами и документами.

```
CREATE TABLE Postings (  
    document_id INT,  
    word_id INT,  
    position INT,  
    FOREIGN KEY (document_id) REFERENCES Documents(id),  
    FOREIGN KEY (word_id) REFERENCES Words(word_id),  
    PRIMARY KEY (document_id, word_id, position)  
);
```

Таблица Links содержит информацию о ссылках между документами.

```
CREATE TABLE Links (  
    source_document_id INT,  
    target_document_id INT,  
    FOREIGN KEY (source_document_id) REFERENCES Documents(id),
```

```

FOREIGN KEY (target_document_id) REFERENCES Documents(id),
PRIMARY KEY (source_document_id, target_document_id)
);

```

Дополнительные таблицы:

```

CREATE TABLE UserQueries (
    query_id INT PRIMARY KEY,
    query_text VARCHAR(255) NOT NULL,
    timestamp TIMESTAMP DEFAULT CURRENT_TIMESTAMP
);

```

Данная таблица Links хранит историю запросов пользователя.

```

CREATE TABLE SearchResults (
    query_id INT,
    document_id INT,
    rank INT,
    FOREIGN KEY (query_id) REFERENCES UserQueries(query_id),
    FOREIGN KEY (document_id) REFERENCES Documents(id),
    PRIMARY KEY (query_id, document_id)
);

```

Таблица SearchResults – результаты поиска.

Данная схема является упрощенной. Реальная версия может включать дополнительные таблицы и индексы, что обеспечивает более высокую производительность и масштабируемость.

Запрос можно представить вектором слов $q = (q_1, q_2, \dots, q_k)$. Здесь k – количество слов в запросе. Модель запроса может включать веса, входящих в запрос слов. Текст запроса поиска документов по слову приведен ниже.

```

SELECT d.url, d.title
FROM Documents d
JOIN Postings p ON d.id = p.document_id
JOIN Words w ON p.word_id = w.word_id
WHERE w.word = 'example'
ORDER BY d.date_added DESC;

```

На рисунке 2 представлена схема поиска по слову.

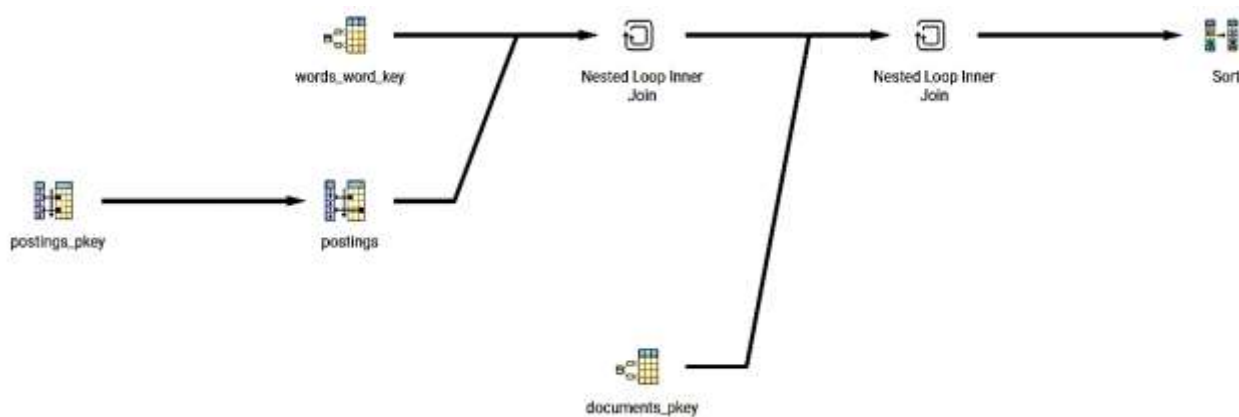


Рисунок 2 – Схема выполнения запроса «Поиск по слову»

Тест запроса на получение связанных документов:

```
SELECT d.url, d.title
FROM Documents d
JOIN Postings p ON d.id = p.document_id
JOIN Words w ON p.word_id = w.word_id
WHERE w.word = 'example'
ORDER BY d.date_added DESC;
```

База данных была создана с использованием системы управления базами данных PostgreSQL. PostgreSQL работает на операционных системах Linux, Windows, macOS и другие Unix-подобные системы. Выбранная нами СУБД не имеет ограничений на использование. Модификацию и распространение кода (лицензия BSD). Выбранная нами СУБД поддерживает множество функций. Это и поддержка сложных типов данных, и полная поддержка транзакций. PostgreSQL имеет встроенные механизмы репликации и отказоустойчивости. Пользовательские функции могут создаваться на различных языках программирования высокого уровня. PostgreSQL предоставляет пользователю мощные средства защиты данных. Данная СУБД включает шифрование соединений и данных на диске. Масштабирование обеспечивает его использование в проектах разного масштаба.

Список литературы

1. Белов, В. В. Повышение пертинентности поиска в современных информационных средах : учебное пособие / В. В. Белов, А. А. Терехов, В. И. Чистякова. – Москва : Горячая линия-Телеком, 2012. — 158 с. — ISBN 978-5-9912-0223-7.

2. Селетков, С. Н. Аналитические исследования в информационных и коммуникационных средах : учебное пособие / С. Н. Селетков, Н. В. Днепровская. – Москва : ЕАОИ, 2010. – 184 с. – ISBN 978-5-374-00368-0.

3. Сидорова, Н. П. Базы данных: практикум по проектированию реляционных баз данных / Н. П. Сидорова. – Москва : Директ-Медиа, 2020. – 92 с. – ISBN 978-5-4499-0799-8

4. Бантюков, С. М. Создание интеллектуальной системы управления качеством предприятия «Эталон» в авиационной промышленности / С. М. Бантюков // Моделирование систем и процессов. – 2024. – Т. 17, № 2. – С. 15-23. – DOI 10.12737/2219-0767-2024-17-2-15-23. – EDN YHXUPY

5. Горбунов, В. Г. Сравнительный анализ методов «Прометей» и нечетких отношений в условиях принятия решений / В. Г. Горбунов, О. Л. Бордюжа, А. А. Пак // Моделирование систем и процессов. – 2024. – Т. 17, № 1. – С. 42-56. – DOI 10.12737/2219-0767-2024-17-1-42-56. – EDN LSMBMD.

References

1. Belov, V. V. Increasing search persistence in modern information environments : a textbook / V. V. Belov, A. A. Terekhov, V. I. Chistyakova. – Moscow : Hotline-Telecom, 2012. 158 p. ISBN 978-5-9912-0223-7.

2. Seletkov, S. N. Analytical research in information and communication environments : a textbook / S. N. Seletkov, N. V. Dene-provskaya. Moscow : EAOI, 2010. 184 p. ISBN 978-5-374-00368-0.

3. Sidorova, N. P. Databases: a workshop on the design of operational databases / N. P. Sidorova. Moscow : Direct-Media, 2020. 92 p. ISBN 978-5-4499-0799-8

4. Bantyukov, S. M. Creation of an intelligent quality management system of the Etalon enterprise in the aviation industry / S. M. Bantyukov // Modeling of systems and processes. – 2024. – Vol. 17, No. 2. – pp. 15-23. – DOI 10.12737/2219-0767-2024-17-2-15-23. – EDN YHXUPY

5. Gorbunov V. G., Bordyuzha O. L., Pak A. A. Comparative analysis of Prometheus methods and unclear relationships in decision-making // Modeling of systems and processes. – 2024. – Vol. 17, No. 1. – pp. 42-56. – DOI 10.12737/2219-0767-2024-17-1-42-56. – EDN LSMBMD.