DOI: 10.58168/MoInSyTe2024_501-505
УДК 004.9

ПРЕИМУЩЕСТВА И НЕДОСТАТКИ ПОИСКОВЫХ БАЗ ДАННЫХ С ОТКРЫТЫМ КОДОМ

Е.В. Чернышова[1], Д.С. Кукуева[1], А.Е. Наденов[1]

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

Аннотация. В статье представлены примеры поисковых баз данных с открытым кодом, их преимущества и недостатки. Также рассматриваются реляционные и нереляционные модели.

Ключевые слова: информационная система, моделирование, база данных.

ADVANTAGES AND DISADVANTAGES OF OPEN SOURCE SEARCH DATABASES

E.V. Chernyshova[1], D.S. Kukueva[1], A.E. Nadenov[1]

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

Abstract. The article presents examples of open source search databases, their advantages and disadvantages. Relational and non-relational models are also considered.

Keywords: information system, modeling, database.

В современном мире нельзя представить функционирование ни одного предприятия без использования баз данных, представляющих собой совокупность структурированной информации, которая обладает определенным стандартом её хранения, манипулирования и редактирования.

Базы данных занимают главную роль при обработке, структуризации, фильтрации и поиске разнообразных совокупностей информации. Есть несколько факторов, отличающих работу с базами данных от работы с другими типами структур данных, например, таблицами [4].

Так, базы данных позволяют комфортно и быстро для пользователя работать с большим объёмом данных, а система запросов является полезным инструментом для фильтрации и агрегации данных. Также стоит выделить возможность работы с сервером в большинстве популярных систем управления базами данных, а значит гарантированную функцию редактирования и просматривания информации из одного источника несколькими пользователями, что необходимо при работе с вэб-приложениями.

Среди баз данных, принято выделять реляционные и нереляционные базы данных. Первые, SQL, характеризуются наличием таблиц и специальных связей между ними. Нереляционные же, noSQL, отходят от табличного представления и созданы специально для определенных типов данных и работы с ними.

Если ранее наиболее популярными и используемыми были реляционные базы данных, то в современном мире больше внимания уделяется работе с нереляционными базами. К ним относятся: базы данных документов, базы данных “ключ-значение”, графовые базы данных и поисковые базы данных.

Поисковые базы данных, или базы данных поисковых систем представляет собой массивы, в которых хранятся данные, собранные модулями для дальнейшего индексирования поисковой системы.[3] Данные модули также называются поисковыми ботами. Поисковые базы данных принято делить на два вида: основной индекс и временная база.

Основной индекс представляет собой хранилище информации, организованное с использованием динамически масштабируемых кластеров. В нем содержатся сокращенные версии веб-документов, включающие ключевые фразы и фрагменты текста, окружающие их, а также ссылки на исходные страницы. Этот подход позволяет значительно ускорить процесс выбора контента, соответствующего введенному запросу, благодаря применению алгоритмов обратного действия, а также уменьшить размер самого индекса.

Временная база содержит в себе результаты индексации ресурсов, где новый контент появляется как минимум один раз в сутки (например, блоги, онлайн СМИ, информационные порталы). Оценка страниц, добавленных в базу, зависит от внутренних факторов оптимизации конкретного документа (таких как соответствие использованных ключевых слов тематике текста, частота их употребления, уникальность). Временный индекс очищается после каждого обновления, а данные из него переносятся в основной. Для оценки качества контента используются стандартные алгоритмы.[1]

Например, не каждый поисковик имеет свои собственные базы данных. Обладают ими лишь крупнейшие участники рынка, такие как Yandex или Google. Другие сервисы используют их наработки. Например, российские Mail.ru и Rambler основаны на алгоритмах и данных, предоставляемых Яндексом, в то время как американский AOL использует базу данных Google. Это связано с необходимостью иметь значительные вычислительные мощности для сбора, хранения и обработки больших объемов информации, на что небольшие компании не могут позволить себе пойти (например, на май 2016 года в базе данных Яндекса насчитывалось более 30 миллиардов веб-документов).[5]

При рассмотрении плюсов и минусов поисковых баз данных с открытым исходным кодом стоит взять за пример базы данных, реализованные с помощью свободной библиотеки Apashe Lucene.

С точки зрения скорости, нет аналога, который мог бы сравниться с Apache Lucene. Это преимущество обусловлено использованием языка программирования Java. Результат запроса занимает всего доли секунды, и это делает его очень эффективным решением для работы любой организации. По мере увеличения скорости растет и общая производительность. Apashe Lucene также имеет небольшое требование к оперативной памяти, максимум 1 МБ. Кроме того, его инкрементная индексация выполняется быстрее, чем пакетная индексация.

В настоящее время оно бесплатно для всех типов использования, включая в том числе и коммерческие цели. По этой причине данное программное обеспечение весьма выгодно для предприятий, которые не обладают большими денежными ресурсами. Также, Apache Lucene предоставляет пользователю полный исходный код, поэтому организации, использующей его, не нужно переписывать свой собственный код.

Важнейшим плюсом поисковых баз данных с открытым кодом является также и то, что благодаря открытому и бесплатному распространению, разработчик всегда может внести свой вклад в улучшение работы программного обеспечения и дальнейшего развития технологии.

Также стоит отметить и минусы.

Из-за открытого исходного кода разработчик может самостоятельно вызвать сбои в работе поисковой базы данных, устранение которых потребует от него наличия большего количества специализированных, углубленных знаний. Также у Lucene есть проблемы с масштабируемостью. Производительность работы может ухудшаться, когда индекс становится больше.

Список литературы

1. Абдуллин А.А. Модели интеллектуальных интерфейсов поисковых информационных систем / А.А. Абдуллин, В.В. Лавлинский, И.А. Земцов– 2019. – Т. 12, № 2. – С. 4.

2. Джуба С. Изучаем PostgreSQL 10 / С. Джуба, А. Волков – 2018. – Т. 15, № 1. – С. 400.

3. Шипилова Е.А., Некрылов Е.Е., Курченкова Т.В. Анализ и моделирование траекторий поведения пользователей онлайн-сервисов с использованием платформы RETENTIONEERING // Моделирование систем и процессов. – 2022. – Т. 15, № 3. – С. 82-93.

4. Новиков Б.А. Основы технологий баз данных / Б.А. Новиков, Е.А. Горшкова – 2019. – Т. 15, № 3. – С. 238.

5. Ревунков Г.И. Проектирование баз данных / Г.И. Ревунков, Н.А. Ковалева, Е.Ю. Силантьева – 2024. – Т. 14, № 2. – С. 49. – DOI:10.12737/2219-0767-2021-14-2-4-12.

6. Sazonova S.A. Strength test of the industrial building's load-bearing structures / Sazonova S.A., Nikolenko S.D., Zyazina T.V., Chernyshova E.V., Kazbanova I.M. – В сборнике: Journal of Physics: Conference Series. III International Conference on Metrological Support of Innovative Technologies (ICMSIT-III-2022). Krasnoyarsk. – 2022. – С. 22016

7. Шипилова Е.А., Платонов А.А., Равлык Р.Ф., Господ А.А. Математическое моделирование и программная реализация процесса управления обеспечением безопасности полетов и деятельностью авиационного персонала // Моделирование систем и процессов. – 2022. – Т. 15, № 2. – С. 100-109

8. Разработка специального программного обеспечения стеганографического скрытия информации в аудиофайлах / Жуматий В.П., Денисенко Д.И., Чернышова Е.В. – Информатика: проблемы, методы, технологии. Материалы XX Международной научно-методической конференции. Под редакцией А.А. Зацаринного, Д.Н. Борисова. – 2020. – С. 1022-1031.

9.Полуэктов А.В., Макаренко Ф.В., Ягодкин А.С. Использование сторонних библиотек при написании программ для обработки статистических данных // Моделирование систем и процессов. – 2022. – Т. 15, № 2. – С. 33-41.

References

1. Abdullin A.A. Models of intelligent interfaces of search information systems / A.A. Abdullin, V.V. Lavlinsky, I.A. Zemtsov– 2019. – Vol. 12, No. 2. – P. 4.

2. Juba S. Studying PostgreSQL 10 / S. Juba, A. Volkov – 2018. – Vol. 15, No. 1. – P. 400.

3. Shipilova E.A., Nekrylov E.E., Kurchenkova T.V. Analysis and modeling of behavior trajectories of users of online services using the RETENTIONEERING platform // Modeling of systems and processes. – 2022. – T. 15, No. 3. – P. 82-93.

4. Novikov B.A. Fundamentals of database technologies / B.A. Novikov, E.A. Gorshkova – 2019. – Vol. 15, No. 3. – p. 238.

5. Revunkov G.I. Database design / G.I. Revunkov, N.A. Kovaleva, E.Yu. Silantieva – 2024. – Vol. 14, No. 2. – p. 49. – DOI:10.12737/2219-0767-2021-14-2-4-12.

6. Sazonova S.A. Strength test of the industrial building's load-bearing structures / Sazonova S.A., Nikolenko S.D., Zyazina T.V., Chernyshova E.V., Kazbanova I.M. – В сборнике: Journal of Physics: Conference Series. III International Conference on Metrological Support of Innovative Technologies (ICMSIT-III-2022). Krasnoyarsk, 2022. – С. 22016

7. Shipilova E.A., Platonov A.A., Ravlyk R.F., Lord A.A. Mathematical modeling and software implementation of the process of managing flight safety and the activities of aviation personnel // Modeling of systems and processes. – 2022. – T. 15, No. 2. – P. 100-109.

8. Lavlinsky, V. V. Modeling of processes and systems / V. V. Lavlinsky, A.S. Yagodkin. - Voronezh, 2017. – 119 p.

9.Poluektov A.V., Makarenko F.V., Yagodkin A.S. The use of third-party libraries when writing programs for processing statistical data // Modeling of systems and processes. - 2022. – Vol. 15, No. 2. – pp. 33-41.