Московский авиационный институт (национальный исследовательский университет)

Факультет информационных технологий и прикладной математики

Кафедра вычислительной математики и программирования

Лабораторная работа №7 по курсу «Информационный поиск»

Студент: В. М. Филиппов
Преподаватель: А. А. Кухтичев
Группа: М8О-410Б

Дата:
Оценка:
Подпись:

Москва, 2025

Лабораторная работа №7 «Сжатие»

В этом задании необходимо применить алгоритмы сжатия к индексу. Исследовать изменения в размерах частей индекса, влияние на скорость индексации и поиска. В отчёте нужно указать:

- Выбранный метод сжатия. Привести побитовую схему хранения данных в индексе.
- Описать причины, по которым был выбран именно этот метод сжатия. Влияние сжатия на размер и скорость прохождения индексу.
- Обосновать, почему поиск после внедрения сжатия работает корректно. Как производилось тестирование?

1 Описание

При выполнении данной лабораторной работы я реализовал два метода сжатия

1. Delta encoding
2. Variable-Length Quantity (VLQ) / Varint

1 Delta encoding

В инвертированных индексах ID документов всегда идут по возрастанию (например: 100, 105, 110...). Если хранить их напрямую, они занимают много места. Если хранить разницы (100, 5, 5...), значения становятся маленькими. В сочетании со вторым методом сильно сэкономит нам память при хранении индекса.

2 Varint

Суть метода заключается в том, чтобы использовать 7 бит каждого байта для хранения данных, а старший бит для указания, есть ли еще продолжение у этого числа.

- Если старший бит равен 1, значит, следующий байт тоже относится к этому числу.
- Если старший бит равен 0, значит, этот байт последний.

При таком кодировании небольшие числа будут занимать меньше места. Также кодирование и декодирование происходит очень быстро (всего несколько битовых сдвигов). Однако если числа всегда большие (например, больше 2^{28}), такой формат может занять даже больше места (5 байт вместо 4), так как на каждые 7 бит данных добавляется 1 служебный бит.

2 Исходный код

Теперь при дампинге или чтении в функциях присутствует флаг zip, который указывает на то, каким образом нам нужно читать или записывать данные.

Реализованы три служебные функции writeVarInt, getVarIntSize, readVarInt. Первая нужна, чтобы записать число как varint, второе чтобы понять, сколько байт это число будет занимать, а третья, чтобы прочитать varint число из памяти. Вторая функция необходима для того, чтобы мы могли посчитать корректно смещение списков вхождения.

Корпус документов из 305135 страниц с размером текста 2661.05 Мб сдамплен за 14.1937 секунд, а итоговый размер получился 376500489 байт (385.9 МБ), что в 3,16 раза меньше, чем несжатый индекс. Это очень существенная экономия памяти. Кроме того, внедрение сжатия увеличило скорость доступа к данным, поскольку кэширование стало более эффективным, в оперативную память не загружается сразу очень много чисел при чтении. Кроме того, сокращаются операции работы с диском). Прочитать с устройства 386 МБ быстрее, чем 1.2 ГБ, даже с учетом времени на распаковку процессором.

Но есть и минусы, например

- Распаковка Varint требует арифметических операций и проверок битов, что даёт дополнительную нагрузку на CPU.
- Невозможность произвольного доступа к документу в списке вхождений.

1 Примеры запросов

Кстати, очень заметно, что поиск выполняется немного дольше, если делать запросы "на холодную". То есть, когда блоки памяти не закэшированы процессором, поиск работает медленнее.

```
Enter query: (manchester & united) | messi
Enter query: (NBA | NFL | NHL) & champion
Tokens searching: [nba, nfl, nhl, champion, ]
Top 10 results
[1] https://en.wikipedia.org/wiki/Major_professional_sports_leagues_in_the_United_States_and_Canada TF-IDF: 87.9164
[2] https://en.wikipedia.org/wiki/List_of_sports_figures_considered_the_greatest TF-IDF: 78.4068
[3] https://en.wikipedia.org/wiki/List_of_Jews_in_sports TF-IDF: 78.1504
[4] https://en.wikipedia.org/wiki/Draft_blunder TF-IDF: 73.9464
[5] https://en.wikipedia.org/wiki/Draft_steal TF-IDF: 73.9464
[6] https://en.wikipedia.org/wiki/Draft_(sports) TF-IDF: 73.9464
[7] https://en.wikipedia.org/wiki/Sports_broadcasting_contracts_in_the_United_States TF-IDF: 69.7583
```

```
[8] https://en.wikipedia.org/wiki/History_of_the_National_Hockey_League_on_television TF-IDF: 67.8845
[9] https://en.wikipedia.org/wiki/History_of_ESPN_on_ABC TF-IDF: 67.1572
[10] https://en.wikipedia.org/wiki/List_of_multi-sport_athletes TF-IDF: 66.54
Query time: 0.0131194 sec
Number of results: 2949 items
```

```
Enter query: (NBA | NFL | NHL) & champion & !football
Tokens searching: [nba, nfl, nhl, champion, footbal, ]
Top 10 results
[1] https://en.wikipedia.org/wiki/NBA_regular_season_records TF-IDF: 53.3538
[2] https://en.wikipedia.org/wiki/List_of_teams_that_have_overcome_3%E2%80%931_series_deficits TF-IDF: 46.7721
[3] https://en.wikipedia.org/wiki/Los_Angeles_Lakers TF-IDF: 46.1695
[4] https://en.wikipedia.org/wiki/Stanley_Cup_playoffs TF-IDF: 43.3573
[5] https://en.wikipedia.org/wiki/Atlanta_Hawks TF-IDF: 42.7092
[6] https://en.wikipedia.org/wiki/Presidents%27_Trophy TF-IDF: 42.547
[7] https://en.wikipedia.org/wiki/Octagon_(mixed_martial_arts) TF-IDF: 40.8788
[8] https://en.wikipedia.org/wiki/Ultimate_Fighting_Championship TF-IDF: 40.8788
[9] https://en.wikipedia.org/wiki/Edmonton_Oilers TF-IDF: 40.751
[10] https://en.wikipedia.org/wiki/List_of_Dancing_with_the_Stars_(American_TV_series)_competitors TF-IDF: 40.6825
Query time: 0.00811675 sec
Number of results: 1036 items
```

```
Enter query: (premier league) & !(championship) & !(league 1)
Tokens searching: [premi, leagu, championship, leagu, 1, ]
Top 10 results
[1] https://theguardian.com/football/2025/dec/17/premier-league-teams-africa-cup-of-nations-sunderland-morocco-chelsea-arsenal-aston-villa TF-IDF: 27.3792
[2] https://theguardian.com/football/2023/nov/17/everton-deducted-10-points-premier-league-guilty-financial-fair-play-breach TF-IDF: 26.753
[3] https://theguardian.com/football/2024/mar/11/premier-league-faces-backlash-after-failing-to-agree-financial-deal-with-efl TF-IDF: 26.4991
[4] https://en.wikipedia.org/wiki/List_of_New_York_metropolitan_area_sports_teams TF-IDF: 26.2424
[5] https://theguardian.com/football/premier-league-2012-13/2013/may/20/all TF-IDF: 25.9398
[6] https://theguardian.com/football/2017/apr/05/premier-league-elite-player-performance-plan TF-IDF: 25.7756
[7] https://theguardian.com/football/2025/aug/18/premier-league-kick-it-out-funding-deal-reduced-concerns-charity-income TF-IDF: 25.7756
[8] https://theguardian.com/football/2023/dec/21/european-super-league-boost-court-of-justice-ruling-uefa-fifa-eu-law TF-IDF: 25.3414
[9] https://theguardian.com/football/2018/apr/19/santi-cazorlas-arsenal-career-not-over-says-hopeful-arsene-wenger TF-IDF: 24.9507
[10] https://theguardian.com/football/blog/2025/jun/10/ange-postecoglou-tottenham-sacked-europa-league-post-truth-philosophy TF-IDF: 24.9428
Query time: 0.0247088 sec
Number of results: 711 items
```

```
Tokens searching: [a, th, ]
Top 10 results
[1] https://en.wikipedia.org/wiki/Wikipedia_talk:Manual_of_Style/Biography/2021_archive TF-IDF: 0.224592
[2] https://en.wikipedia.org/wiki/Tennis_performance_timeline_comparison_(women)_(1884%E2%80%931977) TF-IDF: 0.221983
[3] https://en.wikipedia.org/wiki/Glossary_of_baseball_terms TF-IDF: 0.21911
[4] https://en.wikipedia.org/wiki/Baseball_Annie TF-IDF: 0.21911
[5] https://en.wikipedia.org/wiki/Glossary_of_nautical_terms_(A%E2%80%93L) TF-IDF: 0.204274
[6] https://en.wikipedia.org/wiki/Wikipedia_talk:Manual_of_Style/Dates_and_numbers/Archive_160 TF-IDF: 0.203136
[7] https://en.wikipedia.org/wiki/List_of_Super_Bowl_commercials TF-IDF: 0.201375
[8] https://en.wikipedia.org/wiki/Glossary_of_cue_sports_terms TF-IDF: 0.200915
[9] https://en.wikipedia.org/wiki/Tennis_performance_timeline_comparison_(men) TF-IDF: 0.20053
[10] https://en.wikipedia.org/wiki/List_of_attacks_related_to_secondary_schools TF-IDF: 0.200482
Query time: 0.082748 sec
Number of results: 305064 items
```

3 Выводы

Во время выполнения данной лабораторной работы я узнал про способы сжатия индексов Varint и Delta Encoding. Важно, что эти способы сжатия используются не только при компрессии поискового индекса, но и при разработке БД, а также например в формате protobuf. Под каждым из этих алгоритмов сжатия лежит простая идея, однако в этой простоте и находится элегантность и эффективность решения.

Список литературы

[1] Маннинг, Рагхаван, Шютце *Введение в информационный поиск* — Издательский дом «Вильямс», 2011. Перевод с английского: доктор физ.-мат. наук Д. А. Клюшина — 528 с. (ISBN 978-5-8459-1623-4 (рус.))

[2] Список использованных источников оформлять нужно по ГОСТ Р 7.05-2008