Імовірнісна модель для оптимізації підрахунку унікальних подій у вибірках великого розміру

Виконав студент:
групи КА-45
Ободов Володимир
Науковий керівник:
асистент, Кухарев С.О.

- **Об’єкт дослідження**: статистичні дані про події відвідувань сайту та придбання товарів
- **Предмет дослідження**: застосування методів оцінки кількості унікальних подій для оцінки даних подій
- **Мета**: проаналізувати предмет дослідження та дослідити ефективність застосування алгоритмів.

Підрахунок унікальних подій успішно застосовують у електронній комерції, захисті інформації, медицині, та системному адмініструванні.

Приклади застосування:

- виявлення спаму
- аналіз факторів захворювань
- оптимізація баз даних
- виявлення рекламних оголошень, сторінок сайту, призивів до дії з найбільшою конверсією у результат
- завчасне попередження DDoS атак

Для дослідження була взята вибірка реальних даних з аналітики SaaS сервісу

Вибірка була представлена у вигляді 983 757 записів подій у базі даних. Серед подій були розглянуті такі типи як:

- Придбання товару
- Відвідування певної сторінки сайту
- Створені, але не оплачені закази товарів
- Реакція клієнтів на отримані повідомлення

Структура вибірки

Елементи вибірки мали такі необхідні поля для дослідження як:

- Критерій унікальності
  - *account_id - **ідентифікатор реальної людини яка виконала ту чи іншу дію***
- Основний критерій групування даних
  - *tunnel_id - **ідентифікатор рекламної кампанії***
- Додаткові критерії групування
  - *company_id - **ідентифікатор компанії***
  - *leadgentool_id - **ідентифікатор сторінки***
  - *traffic_source_id - **ідентифікатор джерела трафіку***

Схема вкладеності даних

Вибір хеш функції для найбільш оптимального перетворення

Хеш функції	Рядок		Випадковий UUID		Числа	
	Час	Кіл-ть колізій	Час	Кіл-ть колізій	Час	Кіл-ть колізій
Murmurhash3	145 нс	6	259 нс	5	92 нс	0
FNV-1a	152 нс	4	504 нс	4	80 нс	0
FNV-1	184 нс	1	730 нс	5	92 нс	0
DBJ2a	158 нс	5	443 нс	6	91 нс	0
DJB2	156 нс	7	437 нс	6	93 нс	0
SDBM	148 нс	4	484 нс	6	90 нс	0

Перетворення

1. Створюємо 2^n групп, де n - параметр алгоритму що відповідає за необхідну точність

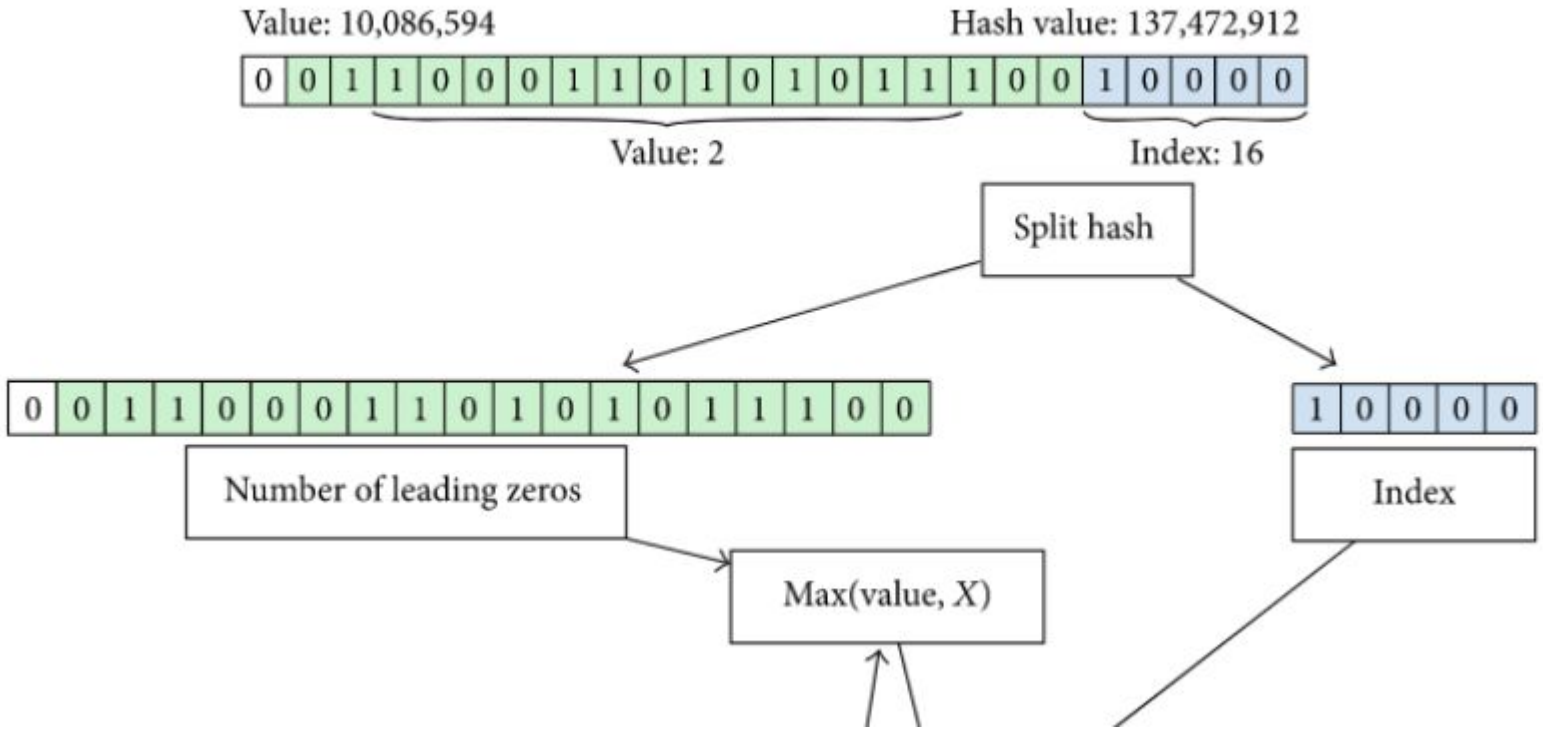
2. Хешуємо кожне значення вибірки, за допомогою функції MurmurHash3, та отримуємо 32 бітне двійкове число

"5b02ca001444c94e4962db2d" => 01001001110101011011010110101

Обробка отриманого хеша

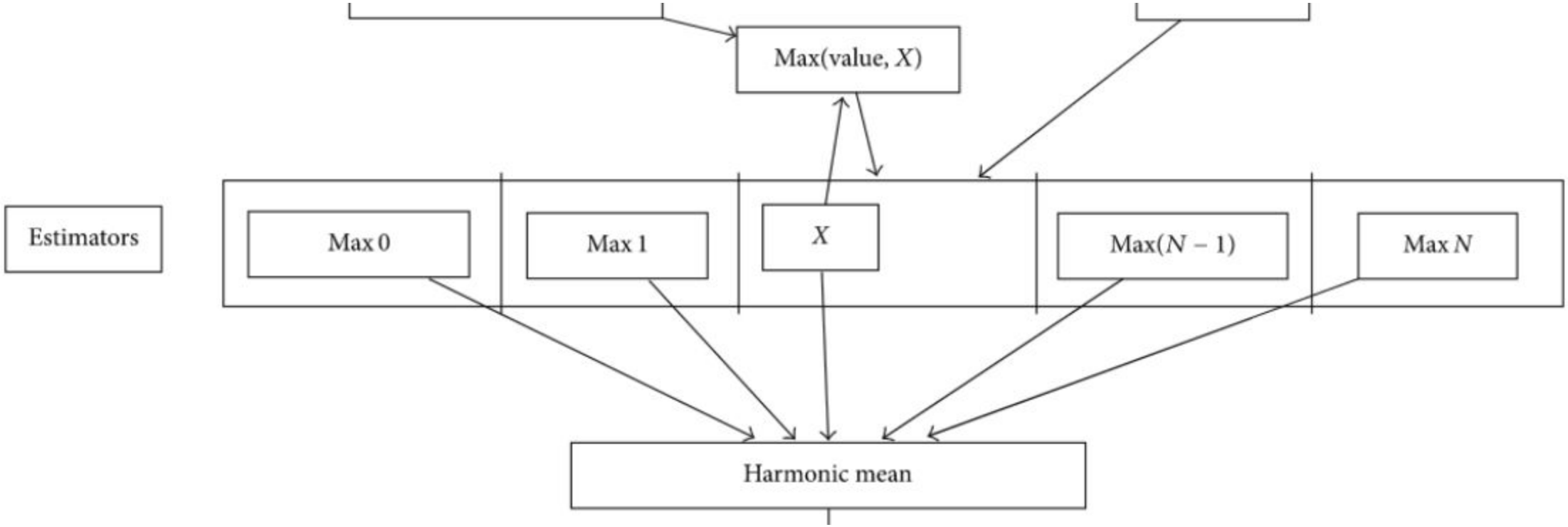
Отримане значення, розподіляється на 2 частини

1. Перші n біт значення використовуються для визначення номеру групи в яку потрапить значення
2. В залишившихся бітах підраховуємо кількість перших нулів, до першої одиниці, для того щоб оцінити імовірність потрапляння цього значення



Оцінка унікальності у групі

1. Кожна з 2^n груп зберігає у собі максимальне число лідуючих нулів
2. Після обробки всіх елементів підраховується гармонічне середнє 2^Mj, де Mj максимальне число в jій группі



Підрахунок кількості унікальних елементів

$$E = a_m m^2 \left(\sum_{j=1}^{m} 2^{-M_j}\right)^{-1}$$

$$\alpha_m := \left(m \int_0^{\infty} \left(\log_2 \left(\frac{2+u}{1+u}\right)\right)^m du\right)^{-1}.$$

Де m - 2^n, Mj - максимальне число перших нулів у j-ій групі

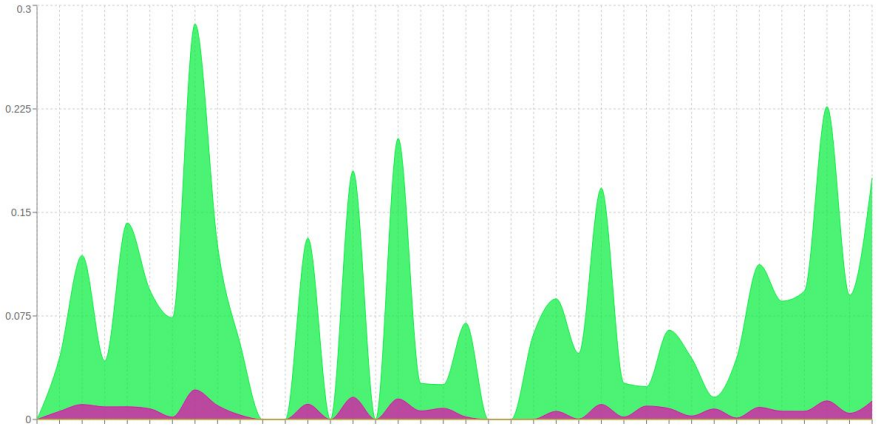
Відхилення методу

Відхилення вираховується за формулою

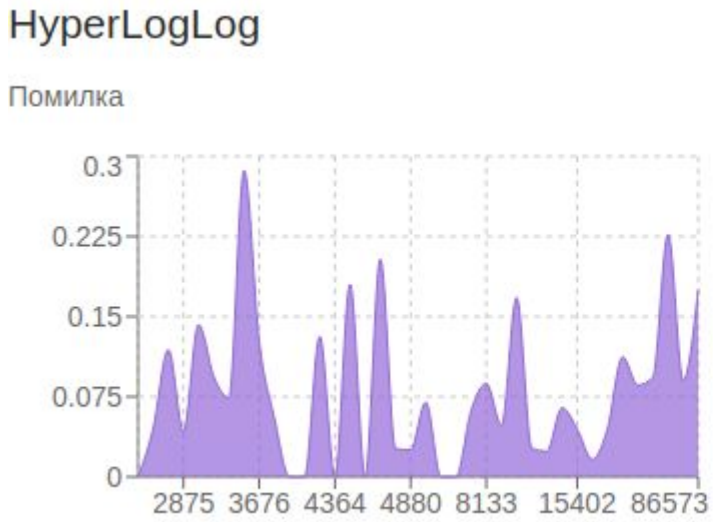
$$\frac{1.04}{\sqrt{m}},\ m = 2^n$$

Тобто n - параметр алгоритму впливає напряму на точність підрахунку. Чим більше n, тим більша точність

Порівняння помилки методів



Метод	Середня Помилка
HyperLoglog	0.0786
HyperLoglog з модифікацією хеш функції	0.0324



Висновки

- Було реалізовано модель з підрахунку унікальних подій у вибірках великого розміру.
- В рамках реалізації було використано алгоритм HyperLoglog та хеш функція Murmurhash3.
- Модифікація хеш функції дозволила покращити результат. Порівняльний аналіз показав що алгоритм HyperLoglog з модифікацією хеш-функції під урахування початкових даних має меншу помилку за класичний алгоритм у середньому в 2.5 рази.

Дякую за увагу

Основні етапи алгоритму HyperLogLog

1. Створюємо 2^n групп, де n - параметр алгоритму що відповідає за необхідну точність
2. Хешуємо кожне значення вибірки, та отримуємо 32 бітне двійкове число
3. Отримане значення, розподіляємо на 2 частини
   a. Перші n біт значення використовуються для визначення в яку группу з 2^n потрапить
   b. В залишившихся бітах підраховуємо кількість перших нулів, до першої одиниці, для того щоб оцінити імовірність потрапляння цього значення
4. Кожна группа зберігає у собі максимальне число лідуючих нулів
5. Після обробки всіх елементів підраховується гармонічне середнє 2^Mj, де Mj максимальне число в jій группі
6. Тоді кількість унікальних подій це E