

HHDB Algorithm

Description of alpha realization

Denis Morozov, PhD

National University of «Kyiv-Mohyla Academy»

19 марта 2014 г.

SQL запрос на выгрузку данных

```
SELECT * FROM access_log
        WHERE lat IS NOT NULL
              AND long IS NOT NULL
              AND homebiz-type != "business"
              AND source = 'lls'
        ORDER BY ip, time
```

Алгоритм построения HouseHolds

Алгоритм построения ip-блоков

```
1:  $i = 0$ 
2:  $cin \gg line, new\ record$ 
3:  $currentIP \leftarrow line.ip, blockCoord \leftarrow line.coord$ 
4: push to resultBase record with  
    $record.ip \leftarrow line.ip, record.coord \leftarrow line.coord, record.time\_begin \leftarrow line.time, record.group\_id \leftarrow i$ 
5:  $i ++, cin \gg line$ 
6: if  $line.ip = currentIP$  then
7:   if  $line.coord \stackrel{CoordAccuracy}{\simeq} currentCoord$  then
8:     goto 6
9:   else
10:    goto 17
11:   end if
12:   storage.clear()
13: else
14:    $record.time\_end \leftarrow previousTime$ 
15:   goto 2
16: end if
17:  $cin \gg line$ 
18: if  $line.time \stackrel{TimeAccuracy}{\simeq} previousTime$  then
19:   storage.push(line)
20:   goto 6
21: else
22:   goto 24
23: end if
24: do the same with storage
```

Алгоритм построения HouseHolds

Пример генерации ip-блоков

```
# block_ip,151.224.40.62
# number of line,10
151.224.40.62,2013-09-16 16:53:13 UTC,,53.42925,-2.128273,BlackBerry,BlackBerry Curve,SK6
151.224.40.62,2013-09-17 16:46:05 UTC,,53.430023,-2.129372,BlackBerry,BlackBerry Curve,SK6
151.224.40.62,2013-09-17 16:47:47 UTC,,53.429276,-2.129783,BlackBerry,BlackBerry Curve,SK6
151.224.40.62,2013-09-17 17:42:11 UTC,,53.42925,-2.128273,BlackBerry,BlackBerry Curve,SK6
151.224.40.62,2013-09-17 18:27:34 UTC,,53.42925,-2.128273,BlackBerry,BlackBerry Curve,SK6
151.224.40.62,2013-09-17 18:58:01 UTC,,53.429585,-2.129502,BlackBerry,BlackBerry Curve,SK6
151.224.40.62,2013-09-18 08:27:51 UTC,c5e90e7622ee6c4846f4a6a7d8edf6b5780299ae,0,0,Android,,n3
151.224.40.62,2013-09-18 08:29:52 UTC,05bdacc10cf0ba4e73e96ff0f669b71a9f440ea3,0,0,Android,,n3
151.224.40.62,2013-09-19 16:54:24 UTC,,53.42861,-2.129557,BlackBerry,BlackBerry Curve,SK6
151.224.40.62,2013-09-19 16:54:26 UTC,,53.42861,-2.129557,BlackBerry,BlackBerry Curve,SK6
# number of id,3

c5e90e7622ee6c4846f4a6a7d8edf6b5780299ae
05bdacc10cf0ba4e73e96ff0f669b71a9f440ea3
# number of coord,6
53.430023,-2.129372
53.429276,-2.129783
53.42925,-2.128273
53.429585,-2.129502
0,0
53.42861,-2.129557
#
```

Идентификация id

По построению идентификация id лога с данным ip происходит по hh-инварианту.

Вероятность id с номером i при привязке данного лога по ip к блоку Block вычисляется по следующей формуле:

$$P(id) = \frac{pr_i(inv_Block)}{\sum_k pr_k(inv_Block)}$$

Замена коэффициента корреляции

Предобработка векторов

Пусть даны два вектора \vec{v}_1 и \vec{v}_2 , представляющие id инварианты. Шаг первый - приведем данную пару к паре векторов \vec{v}_1' и \vec{v}_2' одинаковой длины. Каждый из полученных векторов нормализуем относительно вектора $(1, 1, \dots, 1)$

Замена коэффициента корреляции

Алгоритм нормализации

Для вектора $\vec{v} = (x_1, x_2, \dots, x_n)$ найдем минимум функции

$$f_{(x_1, x_2, \dots, x_n)}(\alpha) = |\alpha x_1 - 1| + \dots + |\alpha x_n - 1|$$

Минимум данной функции будет находиться в одной из точек:

$$\left\{ \frac{1}{x_i} \mid 1 \leq i \leq n \cup x_i \neq 0 \right\}$$

Обозначим его как $\alpha_{min}(\vec{v})$.

Замена коэффициента корреляции

Определим коэффициент отличия двух векторов $\vec{v}_1 = (x_1, \dots, x_n)$ и $\vec{v}_2 = (y_1, \dots, y_n)$ следующим образом

Difference coefficient

$$\text{differenceCoeff}(\vec{v}_1, \vec{v}_2) = \frac{\sum_{i=1}^n |\alpha_{\min}(\vec{v}_1)x_i - \alpha_{\min}(\vec{v}_2)y_i|}{n}$$

Вектора с коэффициентом отличия близким к 0 - похожи.

Общий принцип

На вход подается список координат с временем захвата лога, на выход - результирующая координата.

Общий принцип - исследование сходимости. Важно не общее расположение точек, а центры группировки.

Стабилизация координат ip-блока

Функция $f : \bigotimes_n (\mathbb{R} \otimes \mathbb{R}) \rightarrow \bigotimes_n (\text{list of } \mathbb{N})$

Алгоритм (1-й шаг)

```
1:  $\rho_{min} = \{\min \rho(x_i, x_j) | 0 \leq i, j \leq n\}$ 
2:  $\rho_{max} = \{\max \rho(x_i, x_j) | i, j \leq n\}$ 
3:  $t = 0$ 
4: while  $\rho_{max} - t * \rho_{min} > 0$  do
5:   for ( $i = 1; i \leq n; i++$ ) do
6:      $\alpha_t(x_i) = \text{card}(\{x_j | \rho(x_i, x_j) \leq \rho_{max} - t * \rho_{min}, 0 \leq j \leq n\})$ 
7:   end for
8:    $t++$ 
9: end while
```

Алгоритм (2-й шаг)

После первого шага алгоритма с каждой точкой x_i связана последовательность

$$\alpha(x_i) = (\alpha_1(x_i), \alpha_2(x_i), \dots, \alpha_n(x_i)).$$

Алгоритм (3-й шаг)

Выберем из всех $\alpha(x_i)$ максимальный. Сравнение производится при использовании лексикографического порядка, вначале сравниваем последние координаты. Если максимумов несколько, выбираем самый последний по времени.