HHDB Algorithm

Description of alpha realization

Denis Morozov, PhD

National University of «Kyiv-Mohyla Academy»

15 апреля 2014 г.

Формализация проблемы

Базовыми сущностями положим множество логов. Мотивация для рассмотрения именно такого базового множества, а не множества девайсов, состоит в начальной неопределенности данного множества, а, следовательно, усложнение модели формалистикой теории нечетких множеств.

Степень определенности полей логов

time_stamp	IP	device_id	location
100%	100%	50%	30%

Правильная группировка множества логов позволяет в некоторых случаях получить *device_id* при помощи анализа косвенных данных, предоставляемых множеством логов, методами коллаборативной фильтрации.

Алгоритм постоения HouseHolds

Общая схема алгоритма

- Фильтрация данных
- Сортировка данных
- Создание ip-блоков
- Построение HouseHolds

Дополнительные модули алгоритма

- Стабилизация координат в ip-блоках
- Схема распараллеливания алгоритма

Алгоритм постоения HouseHolds

Фильтрация данных

Данные фильтруются, логи с идентификатором мобильной связи не рассматриваются. В дальнейшем параметры фильтрации могут быть изменены в зависимости от поставленной задачи.

Сортировка данных для построения ip-блоков

Создается скелет для ip-блоков. Для этого из множества логов выбираются логи с определенными координатами из доверительных источников. Полученное множество логов сортируется по ip. В полученных ip-группах данные сортируются по времени.

Алгоритм постоения HouseHolds

Скелет для ip-блока состоит из двух логов - начального и конечного и строится следующим образом.
Каждый следующий лог сравнивается с предыдущим. Если координата отклонилась на величину, меньшую, чем CoordAccuracy, то переходим к следующему логу. Если нет, то проверяем TimeAccuracy.

Сортировка данных

Данный запрос выполняет необходимую для работы алгоритма построения HouseHolds сортировку с частичной фильтрацией:

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
```

Создание ip-блоков

Алгоритм построения ip-блоков включает в себя следующие параметры:

- CoordAccuracy
- TimeAccuracy

Создание ip-блоков

Алгоритм постоения ip-блоков

```
1: i = 0
2: cin ≫ line, new record
3: currentIP ← line.ip, blockCoord ← line.coord
4: push to resultBase record with
   record.ip ← line.ip, record.coord ← line.coord, record.time_begin ← line.time, record.group_id ← i
5: i++, cin ≫ line
6: if line.ip = currentIP then
7:     if line.coord ≃ (CoordAccuracy) currentCoord then
8:         goto 6
9:     else
10:        goto 17
11:    end if
12:    storage.clear()
13: else
14:    record.time_end ← previousTime
15:    goto 2
16: end if
17: cin ≫ line
18: if line.time ≃ (TimeAccuracy) previousTime then
19:    storage.push(line)
20:    goto 6
21: else
22:    goto 24
23: end if
24: do the same with storage
```

Создание ip-блоков

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
#
```

Алгоритм постоения HouseHolds

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 корелляции

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 нормализируем относительно вектора $(1, 1, \ldots, 1)$

Замена коэффициента корелляции

Алгоритм нормализации

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

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

Минимум данной функции будет находится в одной из точек:

$$\{\frac{1}{x_i} | 1 \leqslant i \leqslant n \cup x_i \neq 0\}$$

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

Замена коэффициента корелляции

Определим коэффициент отличия двух векторов $\vec{v_1} = (x_1, \ldots, x_n)$ и $\vec{v_1} = (x_1, \ldots, x_n)$ следующим образом

Difference coefficient

$$differenceCoeff\,(\vec{v_1}, \vec{v_2}) = \frac{\sum_{i=1}^{n} |\alpha_{min}(\vec{v_1})x_i - \alpha_{min}(\vec{v_2})y_i|}{n}$$

Вектора с коэффициентом отличия близким к 0 - похожи.

Стабилизация координат ip-блока

На вход подается список координат с временем захвата лога, на выход - результирующая координата.
Простым естественным методом стабилизации координат ip-блока является взятие центра масс полученных координат.

Метод центра масс

$$v_{block} = \frac{\sum_{i=1}^{n} v_i}{n}$$

Преимуществом данного метода является вычислительная простота. Возращаемая методом координата является хорошим приближением к фактической координате ip-блока при условии высокого процента количества координат логов с незначительным отклонением от фактической координаты ip-блока.

Стабилизация координат ip-блока

Предыдущий метод имеет недостаточную точность, если значительный процент логов имеет неадекватные координаты со значительным отклонением от фактической координаты. В данном случае для нахождение фактической координаты возможно применение следующего принципа.

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 $f : \bigotimes_{n}(\mathbb{R} \bigotimes \mathbb{R}) \to \bigotimes_{n}(\textit{list of } \mathbb{N})$

Алгоритм (1-й шаг)

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

Стабилизация координат ip-блока

Алгоритм (2-й шаг)

После первого шага агоритма с каждой точкой x_i связана последовательность

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

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.