1 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. Правильная же группировка множества логов позволяет в некоторых случаях получить $device\_id$ при помощи анализа косвенных данных, предоставляемых множеством логов методами коллаборативной фильтрации. Введем группировку на множестве логов.

Каждый лог связан с некоторым местом(географическими координатами). Место не всегда фиксируется.

Зададим на множестве логов функцию со следующеей спецификацией (LLW - Logs' Links Weight):

$$koord\_LLW : Logs \times Logs \to \mathbb{R}$$

Данная функция возвращает коэффициент связи между двумя логами по близости географических координат. При высоком коэффициенте связи логи предположительно близки по координатам. Группировка при помощи данной функции предоставляет возможность для разбиения множества логов на кластеры, которые далее могут быть использованы для коллаборативной фильтрации множества логов с условием близости географического положения места захвата логов. Один из возможных алгоритмов кластеризации:

Пример 1.0.1. *Фиксируется некоторое базовое множество логов, задающих кластеры. Например базовое множество можно определить, как множество логов с известными координатами, расположенными относительно далеко друг от друга (например близким к административным центрам). Каждая точка(лог) относится к кластеру, который порожден базовой точкой с наибольшей связью для данной точки, в сравнении с другими базовыми точками. Должно быть определено правило соотношения точки с кластером в случае конфликтов(одинаковая коэффициент связи с несколькими базовыми точками).*

Зададим на множестве логов функцию со следующеей спецификацией:

$$hh\_LLW : Logs \times Logs \to \mathbb{R}$$

Данная функция возвращает коэффициент связи между двумя логами по принадлежности к одному $household\_id$.

Замечание 1.0.1. *Очевидно кластеризация, полученная при помощи функции hh_LLW, должна быть близка к некоторому разбиение кластеризации, полученной при помощи функции $koord\_LLW$. То есть последовательное использование данных двух функций задает некоторую 2-уровневую классификацию на множестве логов. Однако использование различных подходов при реализации данных функций может привести к существенному отклонению от данного правила.*

2 Реализация

2.1 Построение функции $koord\_LLW$

Общая схема - возможна группировка по четким полям при помощи системы запросов к базе данных. Определить систему запросов, позволяющих вычислить значения функции $koord\_LLW$ на области определения.

2.2 Построение функции hh_LLW

Общая схема - аналогична $koord\_LLW$