Кластеризация BigData на примере подарков в ОК

Артур Кадурин (Mail.ru)

highload++ 2014

Конференция разработчиков высоконагруженных систем

«Игрушечный» датасет

- 50.000.000 дарений
- 100.000 разных подарков

Задачи?

- Ранжирование
- Тегирование
- Фильтрация
- Деньги же, ну?..

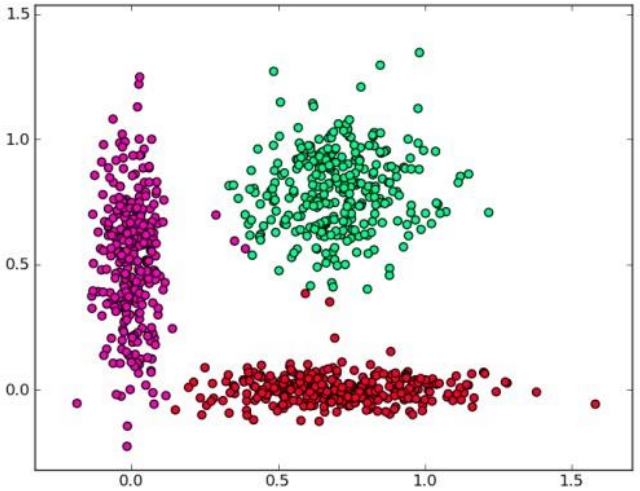
Что такое дарение?

- Даритель: ID, интересы, друзья
- Подарок: ID, теги, картинка, цена
- Получатель: ID, интересы, друзья
- Timestamp: ID праздника, день недели, время суток

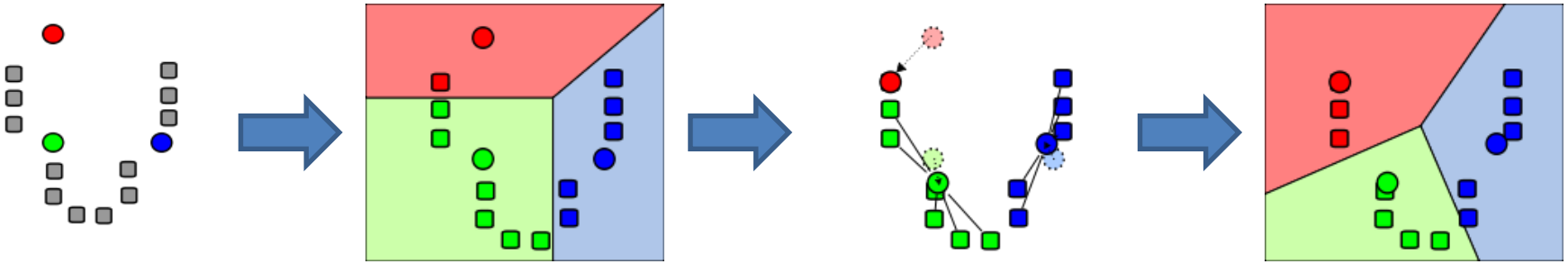
ПвП

Кластеризация

процедура упорядочивания объектов в сравнительно однородные группы



K-Means



Хьюстон, у нас проблема!

- Нет пространства
- Нет расстояний
- Вообще ничего нет
- Варианты?

Основная мысль

Если пользователь подарил два подарка
значит они чем-то похожи

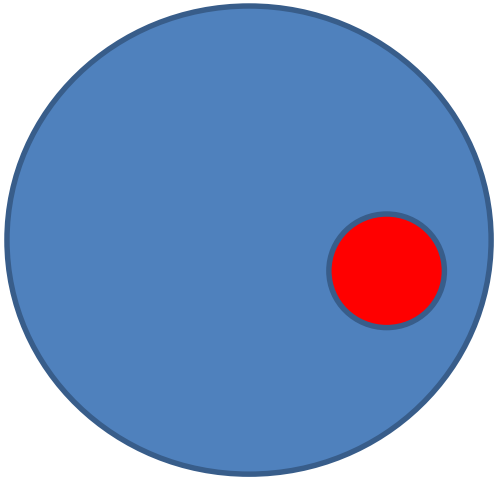
Похожесть..?

$$K(A,B) = \frac{n(A \cap B)}{n(A \cup B)} = \frac{n(A \cap B)}{n(A) + n(B) - n(A \cap B)}$$

$$\frac{\text{количество общих пользователей}}{\text{общее количество пользователей}}$$

Коэффициент Жаккара

Размер имеет значение



Треугольник наоборот

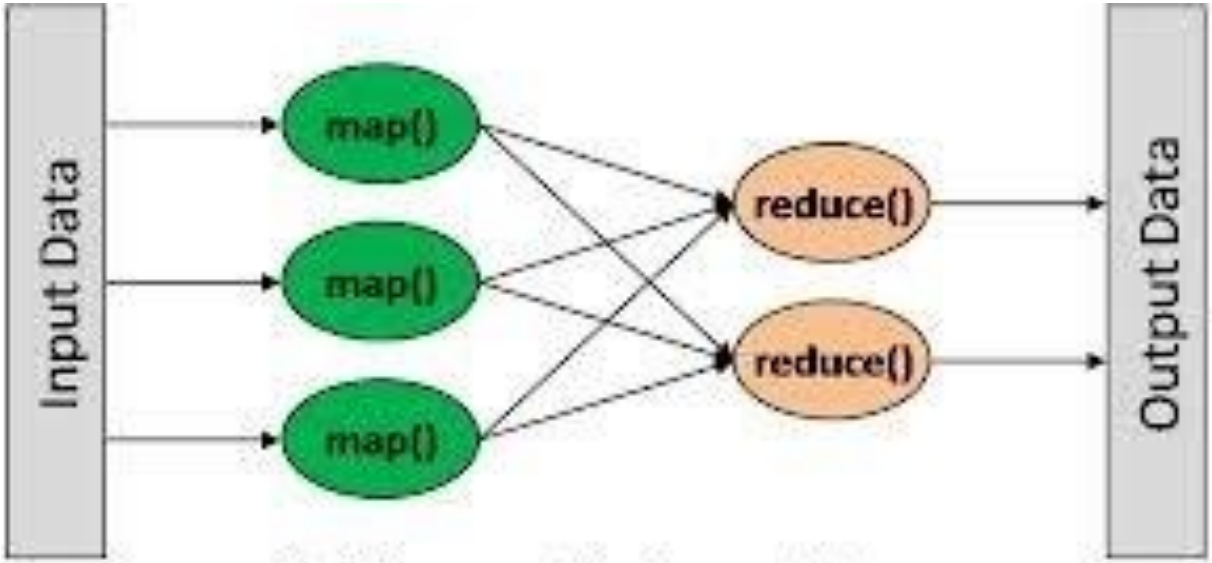
$$1 - K(A, B) + 1 - K(B, C) \geq 1 - K(A, C)$$

1. Lipkus, Alan H (1999), *A proof of the triangle inequality for the Tanimoto distance*, *J Math Chem*
2. Levandowsky, Michael; Winter, David (1971), *Distance between sets*, *Nature*

План

- Есть лог пар Пользователь-Подарок
- Для каждой пары подарков считаем коэффициент Жаккара
- Кластеризуем
- ?????????
- PROFIT

Считаем «похожесть»

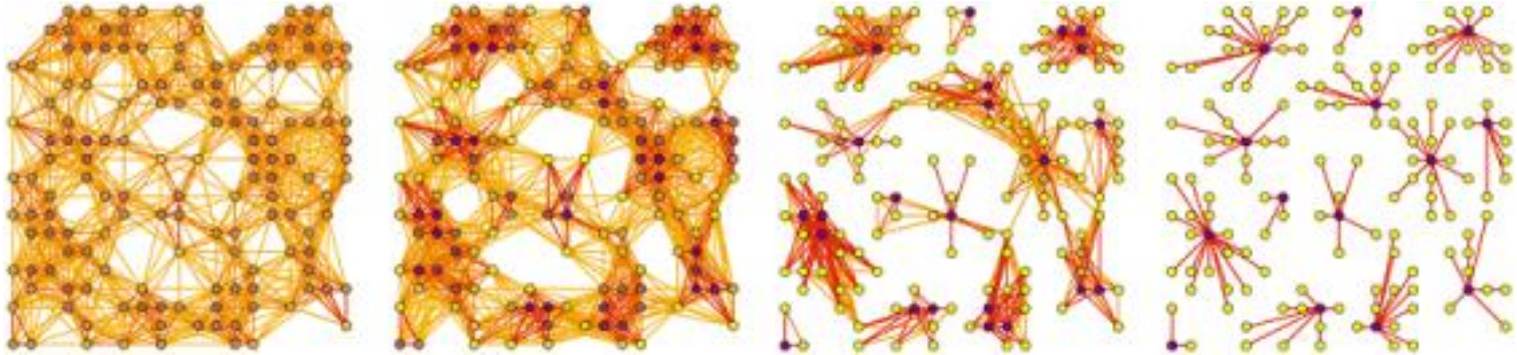


Граф и его матрица

$$A = \begin{bmatrix} 1 & 0{,}209 & \cdots & 0{,}001 \\ 0{,}209 & 1 & \cdots & 0{,}035 \\ \vdots & \vdots & \ddots & \vdots \\ 0{,}001 & 0{,}035 & \cdots & 1 \end{bmatrix}$$

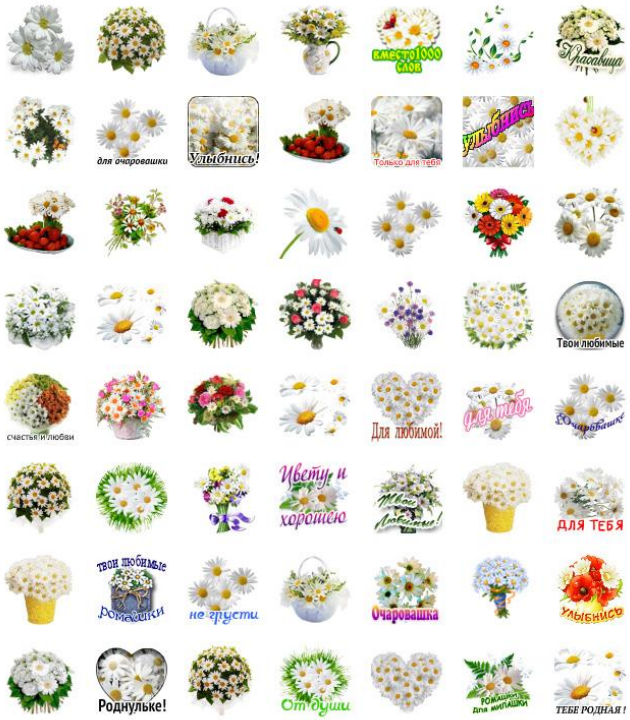
$$B_{ij} = \sum_k A_{ik} A_{kj}$$

$$(\Gamma_r A)_{ij} = (A_{ij})^r / \sum_k (A_{kj})^r$$

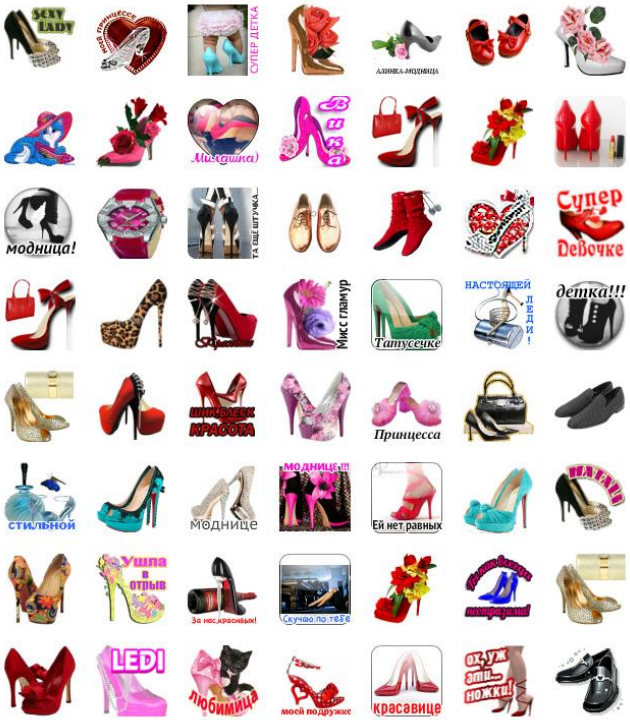


А дальше картинки

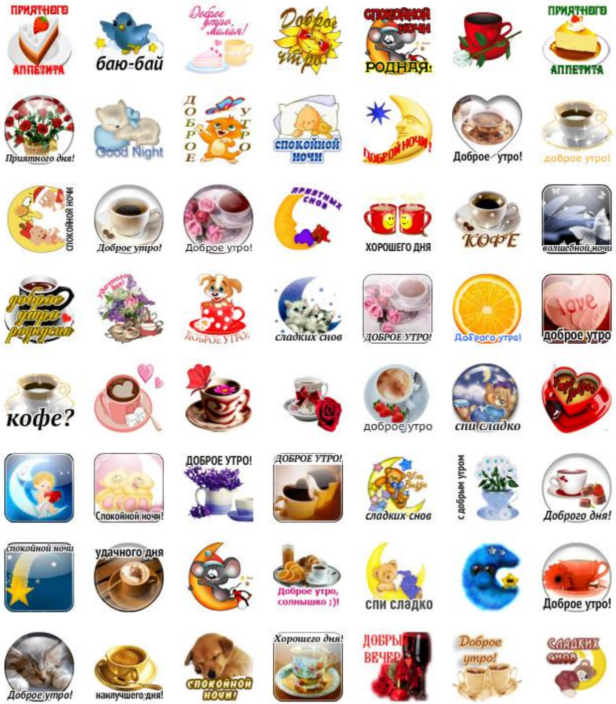
Цветы бывают разные



Яйца и туфли



С новым годом, доброе утро



Женщины и дети

Загнутая подпись

Эксперименты

- Другие расстояния

  по визуальным признакам, с учетом интервала между дарениями, с учетом соц.графа…

- Другая кластеризация

  своя реализация, шанс прохода, «выпихивание»…

- Другие данные

  домены, запросы, товары, туристические направления…

Кластеризация BigData на примере подарков в ОК

Артур Кадурин (Mail.ru)

highload++ 2014

Конференция разработчиков высоконагруженных систем