LATENT VARIABLES IN STATISTICS AND THEIR APPLICATION TO A PROBLEM OF RATING ESTIMATION

Experimental study

A.M. Andronov, A. A.Ressin and K. Pilyushonoka

30 мая 2014 г.

Data volume

Total in Netflix Prize(1) Dataset:

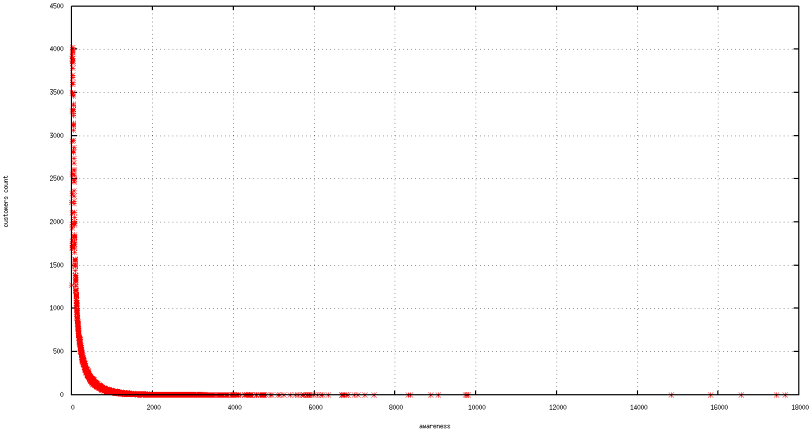
- movies – 17,770;
- customers – 480,189;
- cross-scores – 100,480,507.

Data volume analysis

- Histogram of customer awareness of movies
  - X – user awareness: count of different movies watched by a single customer;
  - Y – count of users with given awareness.
- Histogram of movies popularities
  - X – movie popularity: count of different customers that watched a single movie;
  - Y – count of movies with given popularity.

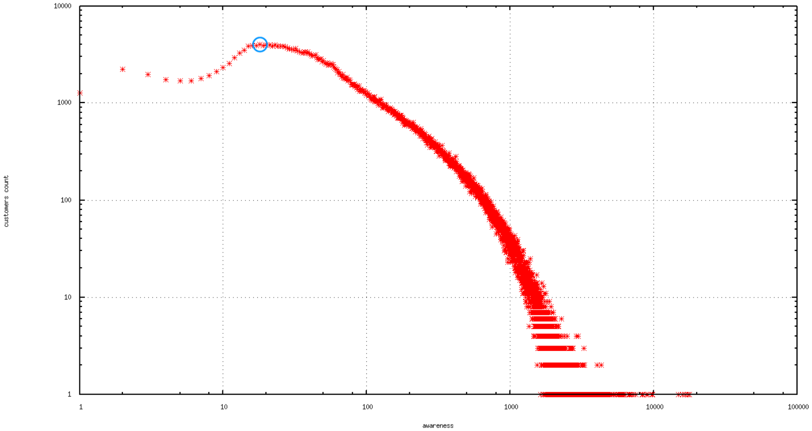
Data volume analysis

Users awareness



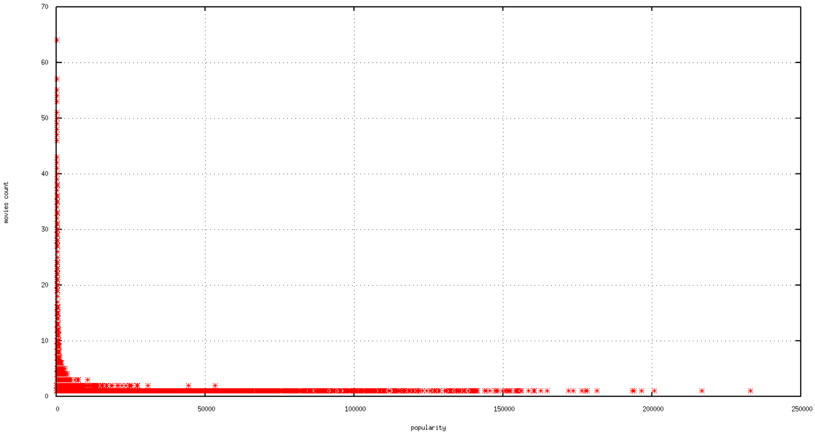
Data volume analysis

Users awareness (log-scale)



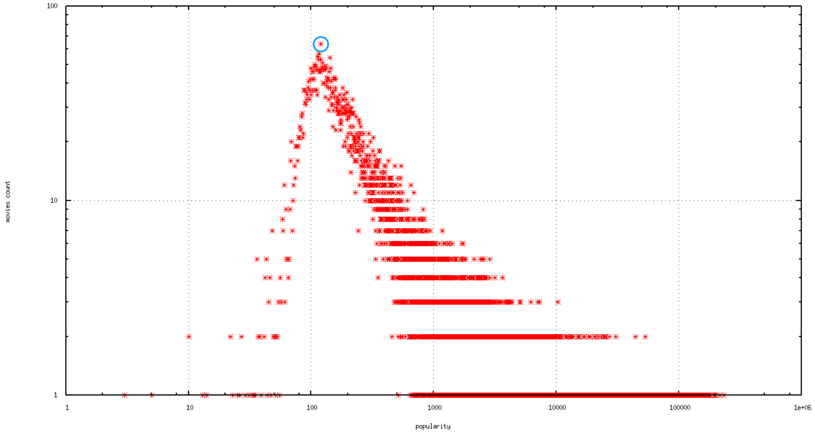
Data volume analysis

Movies popularity



Data volume analysis

Movies popularity (log-scale)



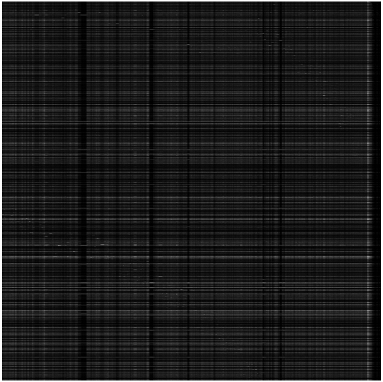
Data volume analysis

Summary

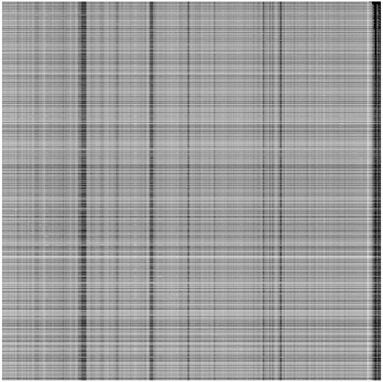
- Awareness:
  - the most frequent (4014 cases) awareness for user is 18;
  - average – 209.25;
  - standard deviation – 302.33.
- Popularity:
  - the most frequent (64 cases) popularity for movie is 120;
  - average – 5654.50;
  - standard deviation – 16909.67.
- Score fill ratio: 1.18%.

Data presence

Sparse matrix visual representation: grayscale



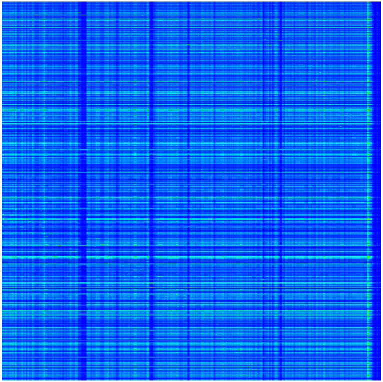
usual



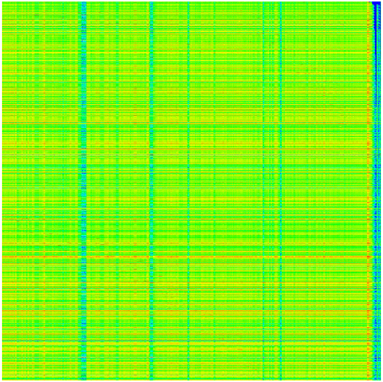
log-scale

Data presence

Sparse matrix visual representation: rainbow



usual



log-scale

Solid submatrix of given sparse matrix

- Let I and J are set of customers and objects indices respectively. The task is to find such $I^* \subseteq I$ and $J^* \subseteq J$, that $\forall i \in I^* \forall j \in J^* \exists y_{ij}$, where y_{ij} is score of object j given by user i and $|I^*||J^*| \to max$.
- The problem is NP-complete, because it is form of famous Clique Problem.
- Need for problem relaxation: maximize submatrix density for given submatrix size $|I^*| \times |J^*|$:

$$|I^*||J^*| - |\{y_{ij} : i \in I^*, j \in J^*, \exists y_{ij}\}| \to min.$$

- The problem still is NP-complete, but now it has heuristical solution.

Heuristical solution

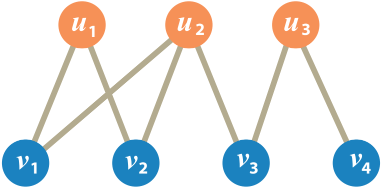
- Let $\mathbf{Q} = \{q_{ij}\}_{n \times m}$ – binary matrix representing data availability.
- Let introduce ϕ_i – authority factors of users, and τ_j – authority factors of movies, such that

$$\phi_i = \frac{\sum\limits_{j=1}^{m} \tau_j q_{ij}}{\sum\limits_{j=1}^{m} q_{ij}}, \text{ and } \tau_j = \frac{\sum\limits_{i=1}^{n} \phi_i q_{ij}}{\sum\limits_{j=1}^{m} q_{ij}}.$$

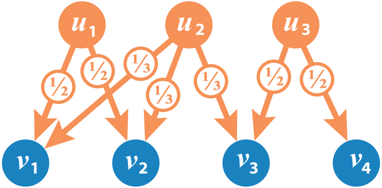
- Take I^* as first $\hat{n}$ indices i from I with maximal ϕ_i, and J^* as first $\hat{m}$ indices j from J with maximal τ_j, i.e.

$$I^* = \{i_1, \dots, i_{\hat{n}}\} \text{ where } \phi_{i_1} > \dots > \phi_{i_{\hat{n}}} > \dots > \phi_{i_n},$$

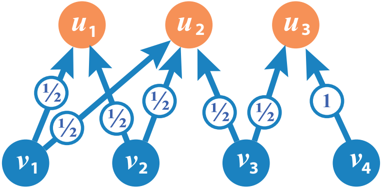
$$J^* = \{j_1, \dots, j_{\hat{n}}\} \text{ where } \tau j_1 > \dots > \tau j_{\hat{n}} > \dots > \tau j_n /$$



$$\mathbf{Q} = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 \end{pmatrix}$$



$$\mathbf{Q}_1 = \begin{pmatrix} 1/2 & 1/2 & 0 & 0 \\ 1/3 & 1/3 & 1/3 & 0 \\ 0 & 0 & 1/2 & 1/2 \end{pmatrix}$$



$$\mathbf{Q}_2 = \begin{pmatrix} 1/2 & 1/2 & 0 \\ 1/2 & 1/2 & 0 \\ 0 & 1/2 & 1/2 \\ 0 & 0 & 1 \end{pmatrix}$$

[1] J. Bennett, S. Lanning, and N. Netflix, "The netflix prize," in *In KDD Cup and Workshop in conjunction with KDD*, 2007.