Factor and Geolocation Based Recommendation

Róbert Pálovics

Informatics Laboratory, Institute for Computer Science and Control , Hungarian Academy of Sciences

Supported by the EC FET Open project "New tools and algorithms for directed network analysis" (NADINE No 288956)
Collaboration between Partners 1 & 3

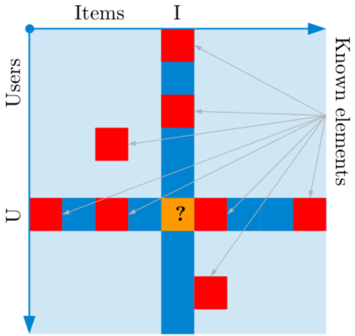
2013

Outline

- Recommender systems
- Collaborative Filtering (CF) methods
- Geolocation related dataset
- CF vs. Geolocation data

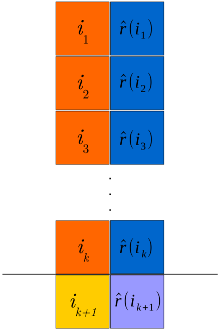
Recommender systems

- Predict the 'rating' or 'preference' that user would give to an item ($\hat{r}$)
- i.e. predict the unknown elements of a user-item matrix



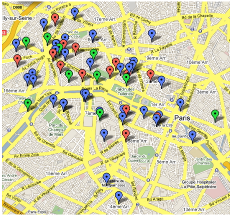
RECOMMENDER SYSTEMS

- Top-k recommendation task: retrieve the best k items for a given user u
  1. Compute $\hat{r}_{ui}$ for all (unknown) items
  2. Order the items
  3. Return the top-k elements in the list



Datasets

- Two datasets, one for France, one for Paris
- User-item scores: how a given user rated a given item
- Item locations: GPS coordinates of the rated items (!)



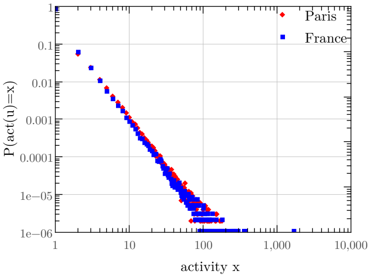
nomao

Basic attributes

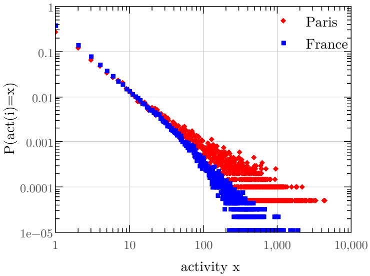
	Paris	France
Number of ratings	1,539,964	1,432,601
Number of users	998,127	1,077,568
Number of items	20,576	99,976
Average ratings per user	1.543	1.329
Average ratings per item	74.84	14.32
Ratio of known ratings	0.0075%	0.0013%

Table: Attributes of the original Paris dataset.

User and item activity



(a) PDF of user activity



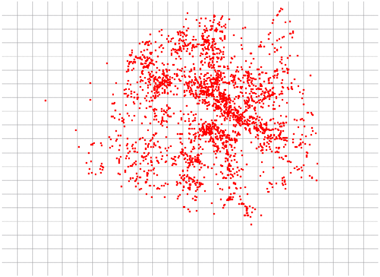
(b) PDF of item activity

Cleaned datasets

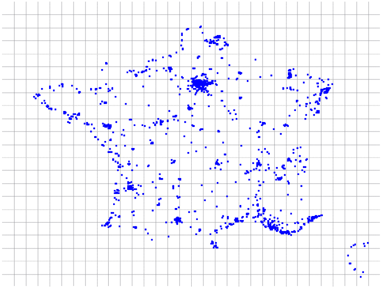
- We select users and items that have at least A ratings between each other
- For Paris we set $A = 10$, for France we set $A = 5$

	Paris	France
Number of ratings	114,352	97,452
Number of users	5,756	9,471
Number of items	2,952	7,605
Average ratings per user	19.87	10.29
Average ratings per item	38.74	12.81
Ratio of known ratings	0.672%	0.135%
Average of rating	3.714	3.747

Map of locations



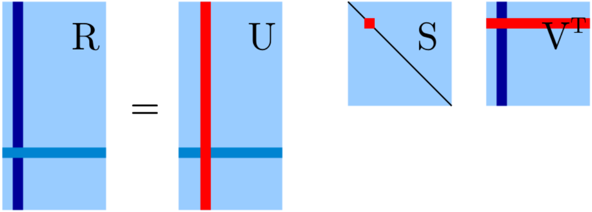
(c) Paris



(d) France

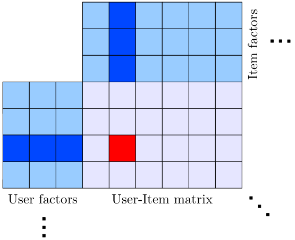
SVD

- $R = USV^T = \sum_{i=1}^{n} u_i \sigma_i v_i$
- $\hat{R}_k = \sum_{i=1}^{k} u_i \sigma_i v_i$
- $\min ||R - \hat{R}|| : \text{rank } \hat{R} = k$
- $\hat{R}_k$ is the best k-rank approximation of the matrix R.



Stochastic gradient descent (SGD)

- Collaborative filtering based recommenders became popular during the Netflix Prize competition[1]
- Large matrix with many unknown values
- $\hat{r}_{ui} = \underline{p}_u \cdot \underline{q}_i + \ldots$
- $\min \sum_{(u,i) \in \text{Train}} |r_{ui} - \hat{r}_{ui}|^2 + \ldots$
- Optimize using SGD

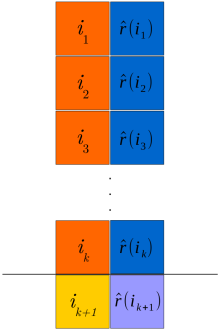


[1] R. Bell and Y. Koren, "Lessons from the Netflix prize challenge," 2007.

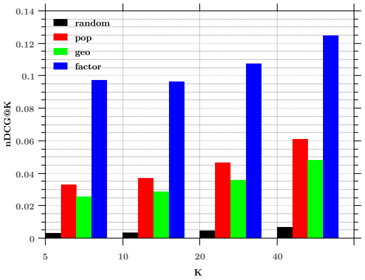
[1] "Netflix update: Try this at home http://sifter.org/∼simon/journal/20061211.html," 2006

Recommender evaluation

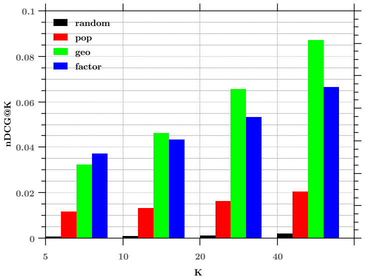
- Remember: top-k recommendation
- Random train + test sets
- Performance measure: *nDCG@K*
- $\text{DCG} = \sum_{i \in K} \frac{r(i)}{\log_2(\hat{rank}(i) + 1)}$
- Baseline recommenders: random, popularity, geo



RESULTS

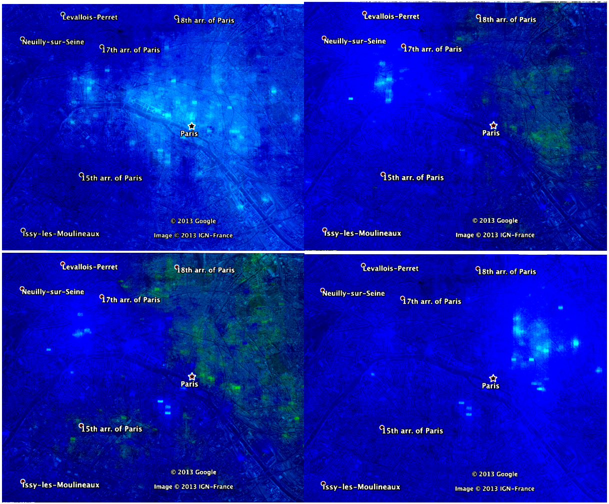


(e) Paris

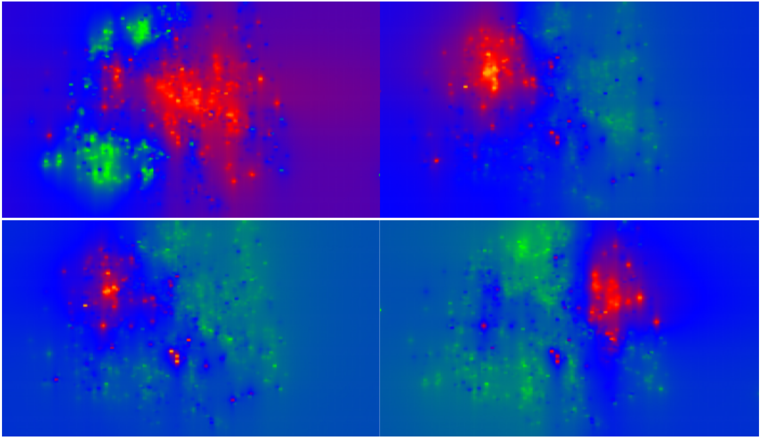


(f) France

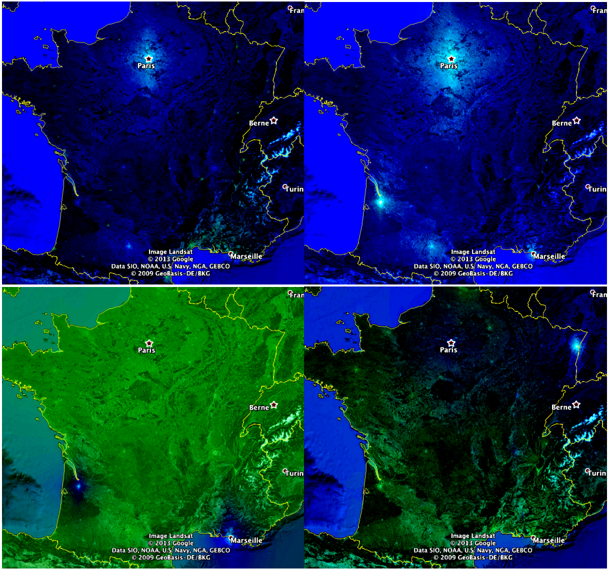
SVD DECOMPOSITIONS - PARIS



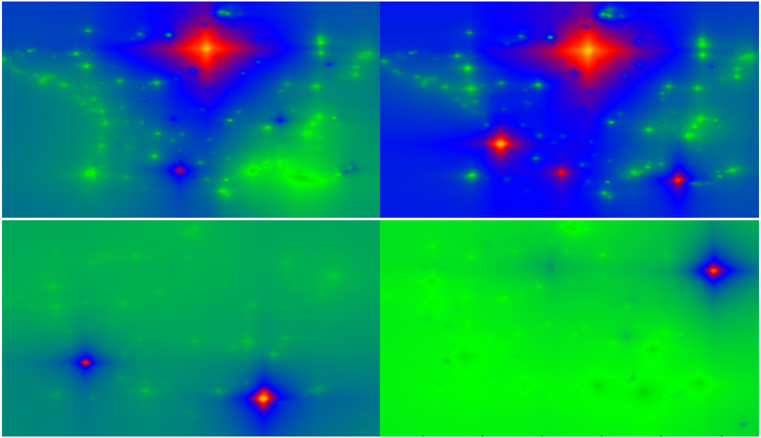
SVD decompositions - Paris



SVD decompositions - France



SVD decompositions - France



Conclusions, future work

- Successful application of the SGD recommender on geolocation based dataset
- SGD can learn geo related features (positive + negative effects)
- Combination, better use of location information
- Create location based networks
- Social regularization (Last.fm):

$$\min \sum_{(u,i)\in \text{Train}} \{|r_{ui} - \hat{r}_{ui}|^2 + \sum_{v\in n(u)} s_{uv}|p_u - p_v|^2 + \sum_{j\in n(i)} s_{ij}|q_i - q_j|^2\}$$