A superparamagnetic agent map of science

University of Zurich[UZH]

Zürcher Hochschule für Angewandte Wissenschaften

Markus Christen (University of Zurich & University of Notre Dame), Thomas Ott (Zurich University of Applied Sciences)

Problem under study

Visualization of high-dimensional data by low-dimensional embedding is an important technique in data analysis. We present a novel solution for this problem based on an agent-based simulation using superparamagnetic clustering and compare it with multidimensional scaling (MDS).

As main application, we present a superparamagnetic agent map (SAM) of data emerging from a survey on the (dis-)similarity of scientific disciplines represented as ISI subject categories. In this way, we map the topology of science using a new paradigm, not relating on citations or keyword-based comparisons of publications, but using the expert opinions of humans.

Superparamagnetic agent maps

Our self-organisation-based heuristic is based on superparamagnetic clustering [1, 2], a nonparametric clustering approach based on local spin interactions. It incorporates the non-metric MDS idea of applying a transformation to the proximities (details in [3]).

A $\mathbf{R}^2$ embedding of a matrix of proximities g_{ij} is constructed by setting up the superparamagnetic clustering framework and performing the following steps:

1. Choose a random point configuation $(\mathbf{x}^0_1; \ldots ; \mathbf{x}^0_n)$ with $\mathbf{x}^0_i \in \mathbf{R}^2$
2. Choose a random spin configuration s^0
3. Set the temperature $T = T_{min}$ and ΔT
4. For T, calculate a new spin configuration s^{t+1} (according to Swendsen-Wang)
5. Calulate the pair correlations $G_{ij}(t + 1) = \delta(s^{t+1}_i, s^{t+1}_j)$
6. For each pair of points, do: If $G_{ij}(t + 1) = 1$ and $J_{ij} > 0$
   then $\mathbf{x}^{t+1}_i = \mathbf{x}^t_i + \alpha\,(\mathbf{x}^t_j - \mathbf{x}^t_i)$
   $\mathbf{x}^{t+1}_j = \mathbf{x}^t_j + \alpha\,(\mathbf{x}^t_i - \mathbf{x}^t_j)$
   else: $\mathbf{x}^{t+1}_i = \mathbf{x}^t_i + \beta \exp(-d^t_{ij})(\mathbf{x}^t_j - \mathbf{x}^t_i)$
   $\mathbf{x}^{t+1}_j = \mathbf{x}^t_j + \beta \exp(-d^t_{ij})\,(\mathbf{x}^t_i - \mathbf{x}^t_j)$
   where $d^t_{ij} = |\,\mathbf{x}^t_i - \mathbf{x}^t_j|$
7. Set $T = T + \Delta T$ and go back to 4 as long as $T < T_{max}$

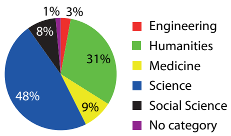
$0 < \alpha < 0.5$ controls the attraction (speed) of two points whose spins are correlated.
$0 < \beta$ controls the repulsion (speed) of two points whose spins are uncorrelated.

Results 1: descriptive statistics of survey data

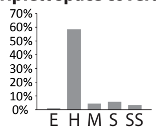
General data:

# of entries total: 785
# of valid entries: 682
# of tripletts: 25'940
tripletts/entry: 38/20
(mean/median)

 # entries per field



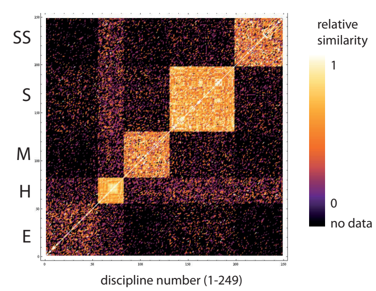
Triplett space coverage



Within/between-field affinity

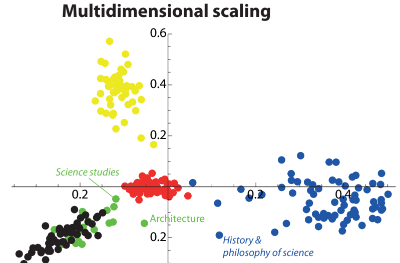
	E	H	M	S	SS
E	0.32	0.12	0.16	0.37	0.18
H	0.12	0.42	0.09	0.13	0.26
M	0.16	0.09	0.37	0.3	0.26
S	0.37	0.13	0.30	0.41	0.21
SS	0.18	0.26	0.26	0.21	0.40

Available distance data



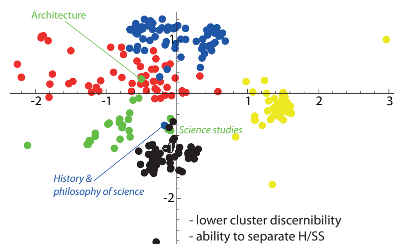
Results 2: Preliminary map of science - MDS versus SAM

Multidimensional scaling



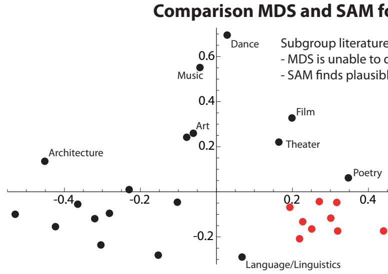
- higher cluster discernibility
- unability to separate H/SS
- more unappropriate topology
  (summarized rank sequence distance: 26)
- unplausibe location of „bridge disciplines"

Superparamagnetic agent map

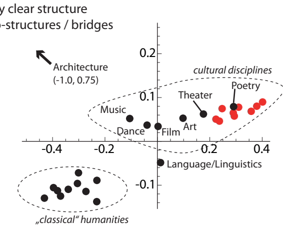


- lower cluster discernibility
- ability to separate H/SS
- more appropriate topology
  (summarized rank sequence distance: 18)
- plausibe location of „bridge disciplines"

Comparison MDS and SAM for Humanities disciplines



Subgroup literature disciplines as plausibility check:
- MDS is unable to display clear structure
- SAM finds plausible sub-structures / bridges



References:
[1] Blatt M, Wiseman S, Domany E. PRL 76, 3251-3254 (1996).
[2] Ott T, Kern A, Steeb W-H, Stoop R. J Stat Mech: theory exp P11014 (2005)
[3] Ott T, Rohrkemper R, Christen M. Superparamagnetic maps: self-organising low-dimensional embedding with superparamagnetic clustering. Working paper.

Survey methodology

We use a a standard procedure of classification psychology that has been translated in a binary comparison task suitable for a web interface. In this data collection procedure, scientists from all disciplines are able to provide their assessments of the similarity of subject category triplets, leading to a distance metrics that integrates a conditional structure (i.e. concept X is attributed to concept Y under the condition that Y has been presented with concept Z).

Choice (1)	Acoustics	Geology
Set point	Mineralogy	
Choice (2)	No similarity	

To manage cominatorical explosion, we made symmetry assumptions (with respect to discipline sequence) and presuppose that disciplines from the same main fields (engineering, humanities, medicine, (social) science) according to ISI are considered to be more similar when compared to a discipline from another field, i.e. subjects that relate themselves to a specific field obtain random triplets where 90% emerge from "their" field. The task is robust for sequence effects and allows that subjects can stop the survey whenever they like. Several distance functions may emerge from the data. For preliminary analysis we use the ratio of positive attributions of two disciplines X and Y compared to the total number of possibilities to attribute X with Y.

Conclusions

Preliminary examination of the data (~26'000 entries) allows to demonstrate that SAM shows superior classification compared to MDS with respect to data topology and plausibility. We find indications that expert-based discipline similarities differ from citation-based similarities. Some presumptions in the survey setup do not hold, requiring a larger data set.

Contribute to a new map of science. Join the survey at www.similarity.ch