HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

MATEMAATTIS-LUONNONTIETEELLINEN TIEDEKUNTA
MATEMATISK-NATURVETENSKAPLIGA FAKULTETEN
FACULTY OF SCIENCE

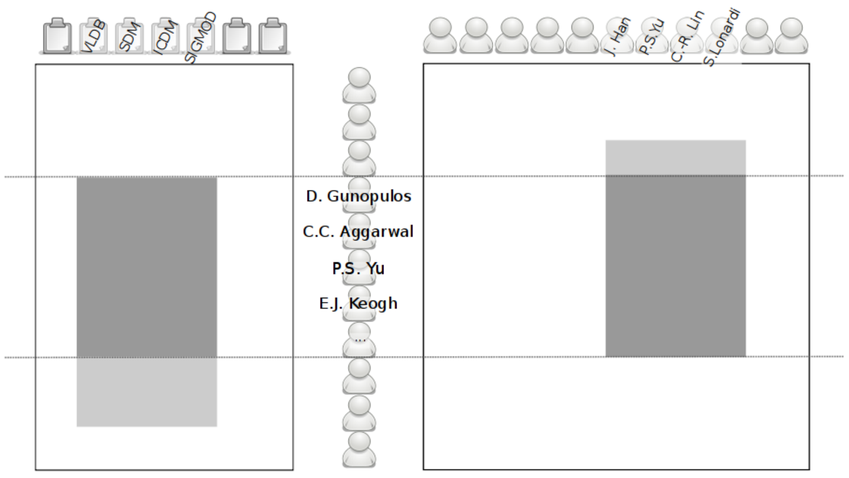
REDESCRIPTION MINING OUTSIDE THE BOOLEAN WORLD

Esther Galbrun and Pauli Miettinen

EXAMPLE

DBLP Bibliography Data

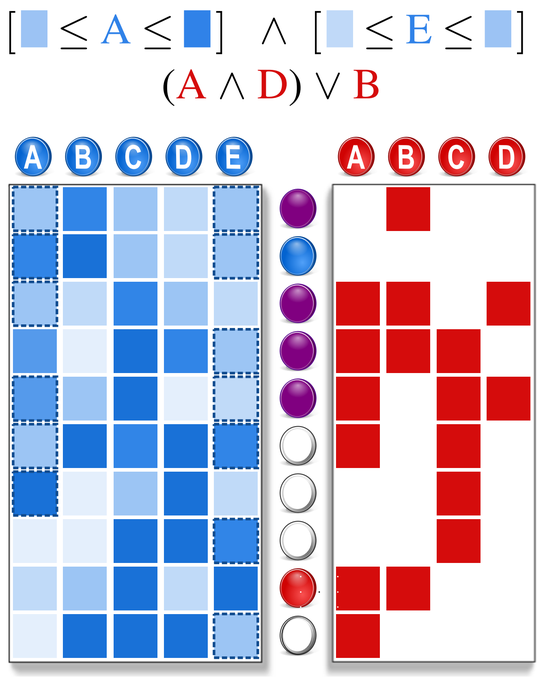
VLDB ∧ ICDM ∧ SDM ∧ SIGMOD
(J. Han ∧ P.S. Yu) ∨ C.-R. Lin ∨ S. Lonardi



OUTSIDE THE BOOLEAN WORLD

Existing methods require **discretization as a pre-processing** step for real-valued data. This can lead to an explosion of the number of variables and generally requires extensive domain knowledge.

This is in contrast to our approach, where **the optimal interval is automatically determined at each step** during the greedy query extension.



FUTURE WORK

Generalizing the niche-finding problem to **traits** and identifying other domains of application (e.g. medicine) is one potential direction for future research.

Improving the algorithm and obtaining proofs of the behavior of redescription mining algorithms for real-valued data are other, more theoretical directions.

PROBLEM DEFINITION

Given two datasets with a bijection between the rows, a **redescription** is a pair of queries $(q_{\mathbf{L}}, q_{\mathbf{R}})$ over the columns characterizing approximately the same sets of rows.
In **redescription mining**, the task is to find the best redescriptions satisfying a given set of constraints.

Approaches to **boolean redescription mining** include decision trees [3], co-clusters [2], frequent itemsets [1] and greedy algorithm [1].

Jaccard coefficient:

$$J = \frac{|E_{1,1}|}{|E_{1,0}| + |E_{1,1}| + |E_{0,1}|}$$

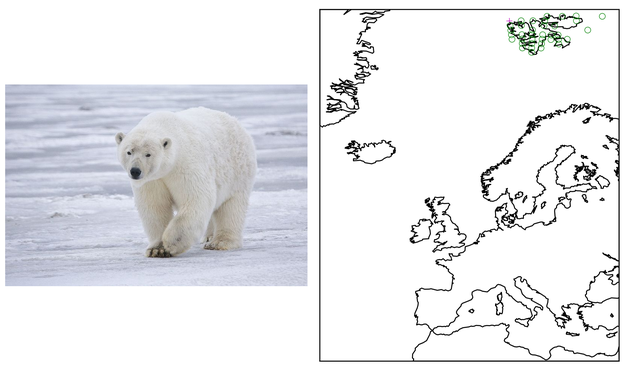
APPLICATION

Bioclimatic Niche Finding

Polar Bear

$[-7.0727 \leq t^{\text{avg}}_{\text{May}} \leq -3.375]$

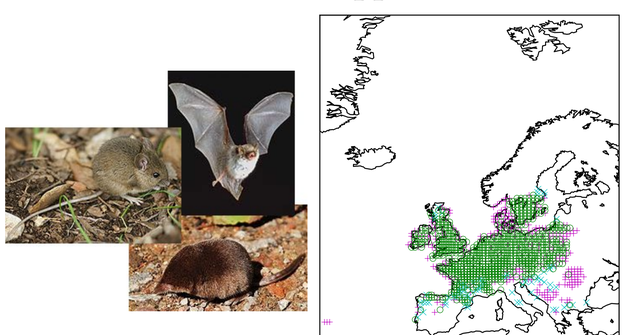
J = 0.973 supp = 36



Wood Mouse ∧ Natterer's Bat
∧ Eurasian Pygmy Shrew

$([3.20 \leq t^{\text{max}}_{\text{Mar}} \leq 14.50] \wedge [17.30 \leq t^{\text{max}}_{\text{Aug}} \leq 25.20]$
$\wedge [14.90 \leq t^{\text{max}}_{\text{Sep}} \leq 22.80]) \vee [19.60 \leq t^{\text{avg}}_{\text{Jul}} \leq 19.956]$

J = 0.623 supp = 681



REFERENCES

[1] Arianna Gallo et al. Finding subgroups having several descriptions: Algorithms for redescription mining. In SDM, pages 334–345, 2008.
[2] Laxmi Parida and Naren Ramakrishnan. Redescription mining: Structure theory and algorithms. In AAAI, pages 837–844, 2005.
[3] Naren Ramakrishnan et al. Turning cartwheels: an alternating algorithm for mining redescriptions. In KDD, pages 266–275, 2004.