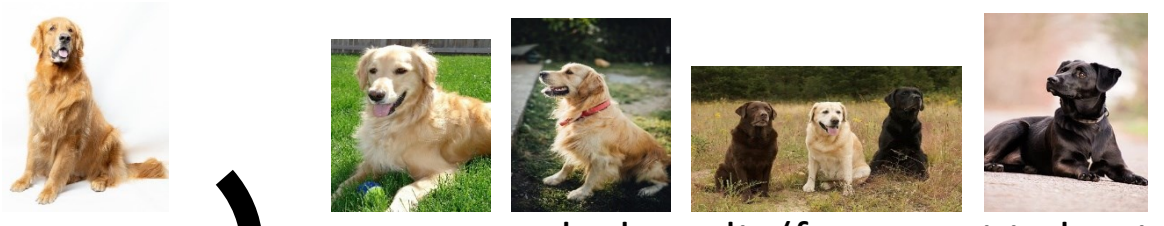
Similarity vs. Relevance: From Simple Searches to Complex Discovery

Tomáš Skopal, David Bernhauer, Jakub Klímek, Petr Škoda, Martin Nečaský
SIRET research group, Faculty of Mathematics and Physics,
Charles University, Prague
www.siret.cz

direct similarity d(.,.) + query-by-example (QbE) + kNN

paradox of the basic model



query → ranked results (from most to least similar)

actually, the best result is the query (example)! d(Q,Q) = 0

→ limited expressive power of direct similarity (effectiveness)

Data-transitive similarity

$$\hat{d}^{\odot,n}_{\biguplus}(\mathbf{x},\mathbf{y}) = \bigodot_{(i_1,\ldots,i_n)\in\mathcal{D}^n} \biguplus \left(d(\mathbf{x},i_1), d(i_1,i_2), \ldots, d(i_n,\mathbf{y})\right)$$

$\mathrm{sum}(\delta_0,\delta_1,\ldots,\delta_n) = \sum_{j=0}^{n}\delta_j$
$\min(\delta_0,\delta_1,\ldots,\delta_n) = \min\{\delta_0,\delta_1,\ldots,\delta_n\}$
$\max(\delta_0,\delta_1,\ldots,\delta_n) = \max\{\delta_0,\delta_1,\ldots,\delta_n\}$
$\mathrm{prod}(\delta_0,\delta_1,\ldots,\delta_n) = \prod_{j=0}^{n}\delta_j$
$\mathrm{iprod}(\delta_0,\delta_1,\ldots,\delta_n) = 1-\prod_{j=0}^{n}(1-\delta_j)$

Examples of inner aggregation $\biguplus$.

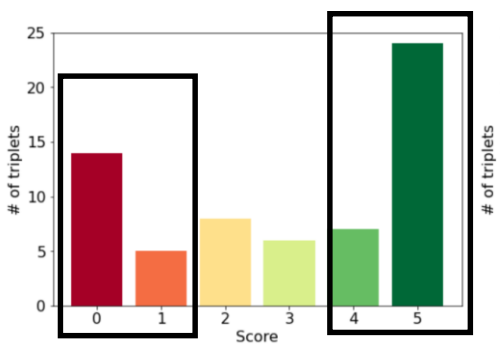
non-metric similarity,
extends relevancy beyond direct similarity

- dataset-dependent (*D*), meta-model: ground distance d(.,.) needed
- aggregation over chains of *n* similar objects from *D*
- self-explainable similarity (the winning chain)

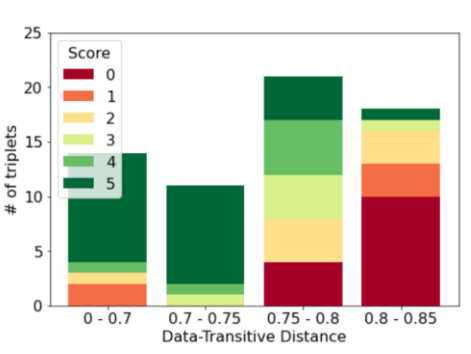
usable in standard QbE searches

Application in dataset discovery

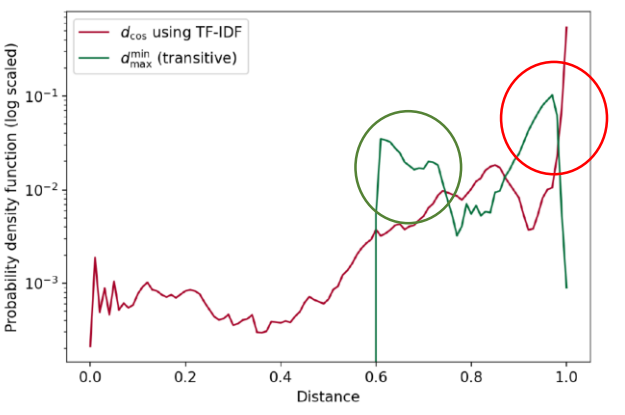
$$\hat{d}^{\min}_{\max}(\mathbf{x},\mathbf{y}) = \min_{\forall i\in\mathcal{D}} \max\{d(\mathbf{x},i), d(i,\mathbf{y})\}$$



78% scores consistent,
57% relevant



complies with
relevance=similarity



	Title Keywords	Description
Q	*Floods in the 19th century* **Floods, Environment, GIS**	Flooded areas in a 19th century flood in the Pilsen region.
I	*5-year water* **GIS, Floods, Environment**	Flooding areas of n-year water in the Pilsen region.
R	*Water reservoirs under the management of the river basin and the forest of the Czech Republic under the territorial jurisdiction of the river Vltava* **water tanks, water management**	The shp file contains points representing water reservoirs whose permitted volume of buoyant or accumulated water exceeds 1 000 000 m3 or to which the Forests of the Czech Republic, p. The registers are updated continuously, the dataset only once a year. The current data can be viewed on the water information portal VODA – www.voda.gov.cz.