Rewriting Queries for Web Searches that Use Local Expressions

Rolf Grütter[1], Iris Helming[1], Simon Speich[1], Abraham Bernstein[2]

[1]Swiss Federal Research Institute WSL, Birmensdorf, Switzerland

[2]University of Zurich, Department of Informatics Zurich, Switzerland

Overview

- Search Queries
- Administrative Regions & Partitions of Regions
- Region Connection Calculus
- Additional Rule “close to”
- Knowledge Base
- Evaluation Methods & Results
- Conclusion & Outlook

Search queries

"Communities close to Zürich"

→ *<u>meaning</u>: all documents regarding the communities close to Zürich*

- vague natural language expression for a spatial relation
- reference point; something is close to something else → providing a hint for scaling

"Close to" in google (p.ex.):

- some public buildings (hotels, offices)
- in cities
- different from meaning of query/ Input

Administrative regions and micro regions

- *Administrative regions*: administrative tasks, institutional structure of a country

- Reflecting a part of humans perception of places and distances between them

---

- *Micro regions*: analysis of spatial mobility, behavior of commuters

- Micro regions in Switzerland: 106 units; a community only part of a single micro region

In Switzerland:

2551 Communities

147 Districts

26 Cantons

Partitions of regions

1. every region has **<u>one</u>** type of partition, e.g. “community” → <u>mutually disjoint</u>

2. <u>Granularity</u> of regions → <u>partitial ordor</u>

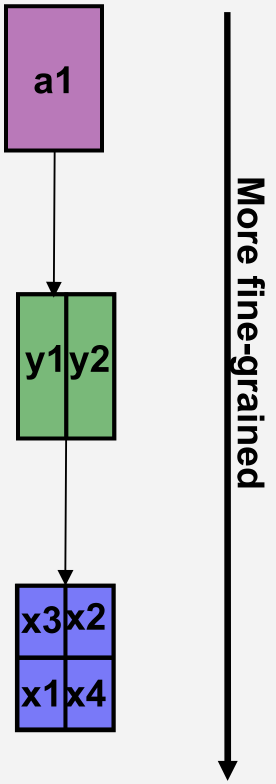
1. cantons
2. districts
3. Communities

$C(x_i)_{i\in I}$ more fine-grained than $D(y_k)_{k\in K}$ $\quad (C(x_i)_{i\in I} \preceq D(y_k)_{k\in K})$ if:

- $(x_i)_{i\in I}, (y_k)_{k\in K}$ partitions of same region (of types C and D)
- each element of $D(y_k)_{k\in K}$ partitioned by elements of $C(x_i)_{i\in I}$

<u>Example:</u> Community$(x_i)_{i\in I}$ and District$(y_k)_{k\in K}$ both partitions of a canton

→ ***reflexive, transitive and antisymmetric***

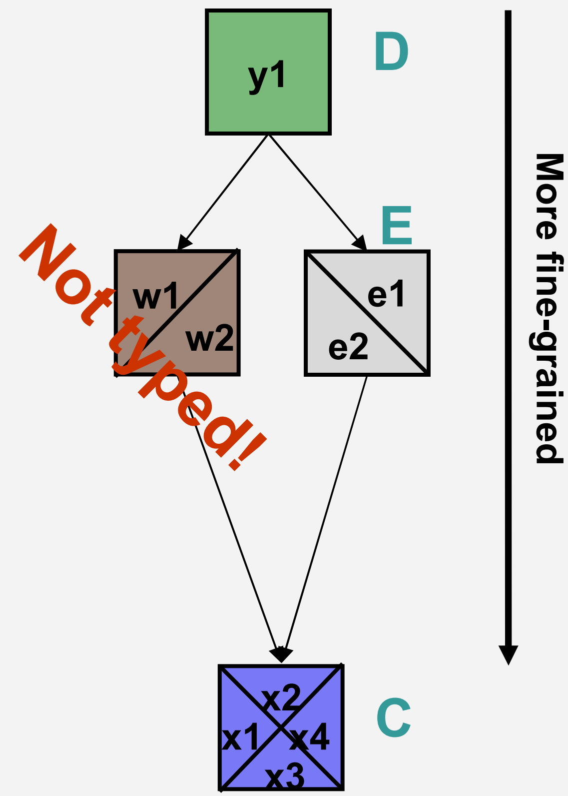


Minimal Partial Order

Regions always have to be typed

Minimal – excluding unwanted partitions

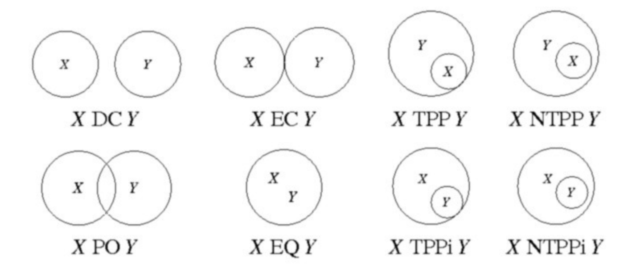
→ *Intransitive*



$C(x_i)_{i \in I} \preceq_{\min} D(y_k)_{k \in K}$ if no type for any $(w_j)_{j \in J}$

such that $C(x_i)_{i \in I} \preceq (w_j)_{j \in J} \preceq D(y_k)_{k \in K}$

Region Connection Calculus with 8 relations (RCC-8)



- Spatial concepts and relations in first order logic
- Spatial or temporal interpretation
- Also other Sets available (e.g. RCC-5)
- Jointly exhaustive and pairwise disjoint
- Formalisms work for few and many data (unlike fuzzy methods)
- Calculating RCC topologies from GIS layers/ spatial databases possible

  (information of administrative regions freely available in many countries)

Additional RCC Relation CLOSE TO

Rule Base: DL-safe SWRL-Rules

Newly added Composition rule:

$$\forall x_a \in (a_i)_{i \in I} \forall y_a \in (a_i)_{i \in I} \forall b \in (b_k)_{k \in K} \forall w$$

$$[\mathrm{P}(x_a, b) \wedge y_a \{\mathrm{P}, \mathrm{EC}\} b \wedge \mathrm{LOC}(x_a, w) \wedge \mathrm{LOC}(y_a, w)$$

$$\rightarrow \mathrm{CL}_{\mathrm{ap}}(y_a, x_a)]$$

Description Logic Knowledge Base

- TBox: partitions as enumerations of individual names (= nominals)
- Linking to types via axioms $C \sqsubseteq \{a_1, \ldots, a_n\}$; where concepts are mutually disjoint $C \sqsubseteq \neg D$
- ABox closed for nominals denoting administrative regions (partOf relations only asserted for partitions)
- {P, EC} represented as auxiliary relation subsuming P and EC

Figures:

- 210 individuals, 21 roles, 12 concepts
- 603 concept assertions
- 29.003 role assertions, e.g.: partOf(Dietlikon, District_Bülach)
- KB grows by square of number of regions asserted; at the moment: canton of Zürich
- Pellet 2.0 as a reasoner

WSL

Evaluation – Searching with “GoForIt”

General Search and Directory Search

Query: ”communities close to Dietlikon”; rewritten: “Nürensdorf OR Dübendorf OR Rümlang OR Wallisellen OR Kloten OR Wangen-Brüttisellen OR Bassersdorf”

- Number of found resources in relevant categories (→ e.g. category: *Nürensdorf*)
  - Recall: related to sum of total resources in same categories
  - Precision: related to all resources in result sets

    - Number of communities resulting from rewriting: between 6 and 24

- 2-sided t-test:
  - Query rewriting significantly increases recall: $p < 0.01$

now: calculating in ≈ 6.3 ms

Results

→ Improvement with query rewriting

Query without rewriting:
"communities close to Dietlikon":

	Total relevant	Total matches	Relevant matches	Recall	Precision
Mean	191.39	14.65	0.10	0.00	--
Max	381	750	1	0.01	1.00
Min	20	0	0	0.00	0.00

Rewritten query: *"Nürensdorf OR Dübendorf OR Rümlang OR Wallisellen OR Kloten OR Wangen-Brüttisellen OR Bassersdorf"* for 170 communities:

	Total relevant	Total matches	Relevant matches	Recall	Precision
Mean	191.39	8,843.50	154.35	0.81	0.07
Max	381	30,880	305	0.91	0.31
Min	20	520	17	0.69	0.00

WSL

Conclusion and Outlook

- Semantics of the relation “close to” may differ in other countries, maybe other formalisms have to be introduced there

- Possible optimizations:
  - Distribution of knowledge bases
  - Outsourcing individuals in database/triple store instead of in-memory storage
  - Move knowledge processing from run-time to design-time

- Using other background knowledge, e.g. travelling time

Additional RCC Relation CLOSE TO

z close to x

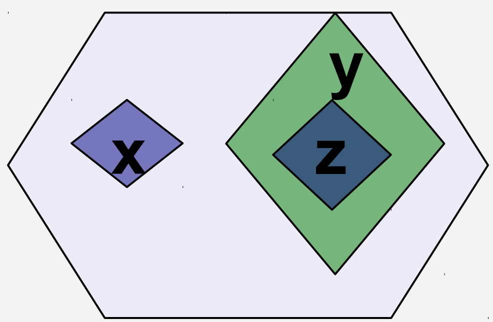
Same partition: symmetrical

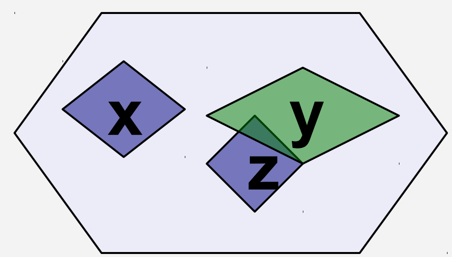
x more fine-grained partition/ z non-administrative region: asymmetrical

→ irreflexive, intransitive (not transitive), not antisymmetric

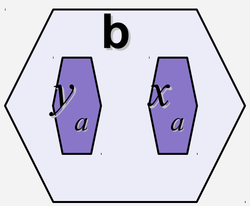
<u>Composition rule 1:</u> $\forall x \forall y \forall z\ [CL_{ap}(y, x) \wedge z\{P, PO\}y \rightarrow CL(z, x)]$

<u>Composition rule 1':</u> $\forall x_a \in (a_i)_{i \in I}\ \forall y_a \in (a_i)_{i \in I}\ \forall z\ [CL_{ap}(y_a, x_a) \wedge z\{P, PO\}y_a \rightarrow CL(z, x_a)]$

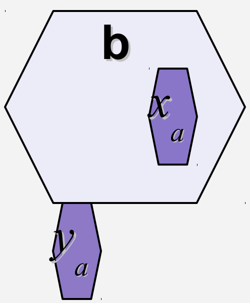




Additional RCC Relation CLOSE TO



Composition rule 2: $\forall x_a \in (a_i)_{i \in I}\ \forall y_a \in (a_i)_{i \in I}\ \forall b \in (b_k)_{k \in K}\ \forall w\ [\mathrm{P}(x_a, b) \wedge$
$y_a \{\mathrm{P}, \mathrm{EC}\} b \wedge \mathrm{LOC}(x_a, w) \wedge \mathrm{LOC}(y_a, w) \rightarrow \mathrm{CL}_{\mathrm{ap}}(y_a, x_a)]$

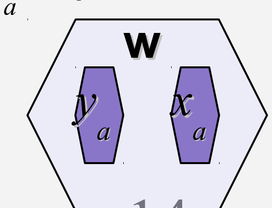


EC: externally connected to

region y_a a priori close to a region x_a, if (i) x_a and y_a same administrative partition $(a_i)_{i \in I}$ (e.g. both communities)

y_a part of or borders same region b of next upper level of administrative partitions $(b_k)_{k \in K}$ (e.g. a district) of which x_a is part

x_a and y_a located (LOC) in the same functional region w



CLOSE TO Algorithm

INPUT: Knowledge Base $\mathcal{KB} = \{\mathcal{T}, \mathcal{A}\}$, Rule Base $\mathcal{RB}$, Concept $\mathcal{Q}$, Individual a

OUTPUT: Set<Individual>

$U \leftarrow \emptyset$, $V \leftarrow \emptyset$, $W \leftarrow \emptyset$, $X \leftarrow \emptyset$, $Y \leftarrow \emptyset$, $Z \leftarrow \emptyset$

$\{b\} \leftarrow \{b \mid \mathcal{A} \models \text{partOf}(a, b)\}$

$U \leftarrow \{u_{i \in I} \mid \mathcal{A} \models \text{partOfOrExternallyConnectedTo}(u_i, b)\}$

$\{c\} \leftarrow \{c \mid \mathcal{A} \models \text{locatedIn}(a, c)\}$

$V \leftarrow \{v_{j \in I} \mid \mathcal{A} \models \text{locatedIn}(v_j, c)\}$

$Y \leftarrow U \cap V$

FOR $(y_k \in Y;\ Y \neq \emptyset;\ Y \setminus y_k)$ {

$X \leftarrow X \cup \{x_{m \in M} \mid \mathcal{A} \models \text{partOfOrPartiallyOverlaps}(x_m, y_k)\}$ }

$W \leftarrow \{w_{n \in N} \mid \mathcal{A} \models Q(w_n)\}$

$Z \leftarrow X \cap W$

OUTPUT Z