Rewriting Queries for Web Searches That Use Local Expressions

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

[1] Swiss Federal Research Institute WSL, Birmensdorf, Switzerland
{rolf.gruetter, iris.helming, simon.speich}@wsl.ch
[2] University of Zurich, Department of Informatics Zurich, Switzerland
{bernstein}@ifi.uzh.ch

Abstract. Users often enter a local expression to constrain a web search to a geographical place. Current search engines' capability to deal with expressions such as "close to" is, however, limited. This paper presents an approach that uses topological background knowledge to rewrite queries containing local expressions in a format better suited to standard search engines. To formalize local expressions, the Region Connection Calculus (RCC) is extended by additional relations, which are related to existing ones by means of composition rules. The approach is applied to web searches for communities in a part of Switzerland which are "close to" a reference place. Results show that query rewriting significantly improves recall of the searches. When dealing with approx. 30,000 role assertions, the time required to rewrite queries is in the range of a few seconds. Ways of dealing with a possible decrease of performance when operating on a larger knowledge base are discussed.

Keywords: Local Expression, Query Rewriting, Region Connection Calculus (RCC), Web Ontology Language (OWL), DL-safe SWRL Rules.

1 Introduction

Web searches are quite often constrained by local references (e.g., place names or expressions for spatial relations between places) [1, 2]. However, existing search engines are weak in supporting spatial queries. While local expressions, such as "close to", may be used as strings in a query, they are usually evaluated according to the frequency of their occurrence in the indexed documents and not according to their meaning in natural language. For some queries, Google[1] returns resources of real world entities which are "close to" or "in the surroundings of" a reference place. However, this feature is limited to government agencies and commercial enterprises such as hotels, surgeries and offices in urban areas, in a way similar to the yellow pages.[2]

[1] http://www.google.ch/
[2] http://yellow.local.ch

In: N. Bassiliades et al. (Eds.) RuleML 2011 – Europe. LNCS, vol. 6826, pp. 345–359. Springer, Berlin Heidelberg (2011)

Localized versions of standard search engines, such as Google, Yahoo[3] and Bing[4], offer the option of displaying query results in the national language or from hosts in the respective country. In addition, there are a number of search engines whose scope is limited to a single country or a geographical region. These engines support queries and return results in the language of the country (e.g., the Chinese search engine Baidu[5]). They also provide local information, such as yellow and white pages. As a motivating example will show, the advantage of a localized search goes beyond "just" finding yellow and white pages in a national language.

This paper presents an approach to support web searches by rewriting queries using topological background knowledge which is created from the administrative structure of a country and from a tessellation of micro regions. The latter establish consistent units for the analysis of spatial mobility. Applying our approach to web searches that use local expressions significantly improves recall. When using an X86-based PC operating on a knowledge base holding about 30,000 role assertions, query rewriting takes 6,317.54 ms on the average.

The paper is organized as follows: Section 2 provides an overview of the Region Connection Calculus (RCC) and DL-safe SWRL rules. RCC is used as a foundation of the formalism which is introduced in section 4. DL-safe SWRL rules are used – together with OWL DL – to implement the formalism which is described in section 5. The same section also shows how queries are rewritten. In section 6, the approach is applied to web searches for communities which are "close to" a reference place and the results of this application are presented. Section 7 discusses related work and section 8 concludes with an outlook on future work.

This paper extends previous work [3] by (i) refining the basic formalism and separating it from background knowledge, (ii) considering spatial mobility as an additional source of knowledge, and (iii) evaluating the approach on the basis of web searches in a realistic scenario.

2 Region Connection Calculus and DL-safe SWRL Rules

The Region Connection Calculus (RCC) is an axiomatization of certain spatial concepts and relations in first order logic [4]. The basic theory assumes just one primitive dyadic relation: C(x, y) read as "x connects with y". Individuals (x, y) can be interpreted as denoting spatial regions. The relation C(x, y) is reflexive and symmetric.

Using the primitive relation C(x, y) a number of intuitively significant relations can be defined. Of these relations, PP ("proper part of"), PPi ("inverse proper part of"), PO ("partially overlaps"), EQ ("equal to") and DR ("discrete from") form a jointly exhaustive and pairwise disjoint set, which is known as RCC-5. Similar sets of one, two, three and eight of these relations are known as RCC-1, RCC-2, RCC-3 and RCC-8, respectively. PP and PPi are subsumed by the relations P ("part of") and Pi ("inverse part of"). RCC also incorporates a constant denoting the universal region, a

[3] http://ch.search.yahoo.com
[4] http://www.bing.com
[5] http://www.baidu.com

sum function and partial functions giving the product of any two overlapping regions and the complement of every region except the universe [4].

According to Randell et al. [4], regions support either a spatial or temporal interpretation. For a spatial interpretation, a topological model is provided. According to this model, regions are interpreted as sets of points in a point-based universe and C(x, y) holds if the topological closures of regions x and y share (at least) a common point. In order to comply with the model-theoretic semantics of Description Logics (DL), the RCC relations are interpreted in this paper as binary relations between individual regions in an abstract domain.

In order to infer new from existing knowledge or to check consistency of a knowledge base holding spatial relations, so-called composition tables are used. The entries in these tables share a uniform inference pattern which can be formalized as composition axioms of the general form $\forall x \forall y \forall z\ [S(x, y) \wedge T(y, z) \rightarrow R_1(x, z) \vee \ldots \vee R_n(x, z)]$ where S, T, and R_i are variables for relation symbols.

RCC composition rules can be implemented as DL-safe SWRL rules. DL-safe SWRL rules are function-free Horn rules with the restriction that each variable in the rule occurs in a non-DL-atom in the rule body [5]. This is ensured by adding special non-DL-literals such as $\mathcal{O}(x)$ to the rule body, and by adding a fact $\mathcal{O}(a)$ for each individual a to the knowledge base.[6] While in theory DL-safe SWRL rules support complex, i.e., disjunctive, heads (or negation in the rule body) [6], there is currently no implementation that supports this feature. However, since the RCC relations are jointly exhaustive [4], it is always possible to replace a negative atom, for instance ¬disconnectedFrom(z, y), by a, possibly auxiliary (cf. section 4), positive atom, for instance connectsWith(z, y).

3 A Motivating Example

Think of a woman taking up a new job in the community of Dietlikon (which is located in the canton of Zurich). She might not be familiar with this part of Switzerland, but still wants to find a home which is close to her place of work. Before calling a housing agency she might want to inform herself about the communities close to Dietlikon by searching the web. The retrieval problem triggered by her information need can be put as follows: "For every community that is close to the community of Dietlikon, retrieve all resources from the web." Note that housing agencies on the web usually offer the opportunity of searching within a selectable Euclidean distance from a reference place. Euclidean distance, however, can be tricky when looking for close places. It does not consider conditions such as topography and local public infrastructure.

To make local expressions such as "close to" meaningful, the approach presented in this paper uses topological background knowledge in terms of spatial relations between administrative units and functional micro regions. Administrative units establish the institutional structure of a country. They are typically organized into a

[6] For the evaluation (cf. section 6) it was sufficient to add a fact $\mathcal{O}(a)$ for each individual a. The requirement that $\mathcal{O}$ must not be a concept from the DL knowledge base was not considered.

set of partially ordered partitions. Units of the same partition share the same type. Each unit is administered by a local authority. Switzerland, for instance, is organized into 26 cantons, 147 districts and 2551 communities [7]. Micro regions, on the other hand, do not contribute to a country's administration. They have been established as consistent units for the analysis of spatial mobility and "encode" things such as the behavior of commuters. In Switzerland, the tessellation of micro regions consists of 106 units [7]. Whereas these form a partition similar to those of administrative units, this does not align with the partial ordering of the latter. However, micro regions still align with the smallest units of institutional organization in that a given community is part of a single micro region only.

It is well documented that administrative boundaries influence how people perceive distance (cf. section 7). Some evidence for this comes from the fact that boundaries, for instance of districts, often take course along natural boundaries such as ridges or watercourses thereby "encoding" some prominent topographic features. Districts further divide a country into units performing decentralized administrative tasks in areas such as health (hospitals), education (schools) and judiciary (courts) [7]. Hence, districts – and administrative units in general – suggest themselves as a foundation for a formalism of proximity. Administrative units, however, do not always properly reflect functional properties such as local public infrastructure. In order to include these, the presented approach also considers a tessellation of functional micro regions. Note that the work presented here is still in progress. Further factors influencing the perception of proximity on different scales of social organization may be added in the future.

Fig. 1. Eight communities close to the community of Dietlikon. Shaded areas show different districts. The bold line borders the functional micro region of Glattal-Furttal.

The formalism introduced in the following section is defined on a topological structure.[7] This is the basic idea: A region z is close to a region x if another region y is *a priori* close to x and z connects with y. Note that the type of x implicitly encodes a scale factor: What is close to a community is not the same as what is close to a district.[8] In the next section, this basic rule is refined and linked to two different sources of background knowledge. In order to get back to the example, our approach evaluates the eight labeled communities in Figure 1 as being close to Dietlikon. They are in the intersection of communities that are part of or externally connected to the district of Bülach and those that are located in the micro region of Glattal-Furttal (bold borderline).

4 A Formalism for Proximity

4.1 The Basic Composition Rule

In order to formalize local expressions, RCC is extended by additional relations. In the context of this paper, $\mathsf{CL}(x, y)$, which is read as "x is close to y", is introduced as a *weakly asymmetrical* relation, in accordance with empirical evidence [9]. Against the background of knowledge considered in this paper, this means that the relation is usually symmetrical, if x and y are members of the same administrative partition (e.g., both are communities), but asymmetrical, if y is a member of a more fine-grained partition than x (e.g., y is a community and x a district) or else, if x is a non-administrative region. $\mathsf{CL}(x, y)$ is further irreflexive, intransitive and not antisymmetric.

The additional RCC relation is related to the existing ones by means of a composition rule in such a way that the rule is a necessary condition for the relation:

Composition rule 1. $\forall x \forall y \forall z\ [\mathsf{CL}_{\mathsf{ap}}(y, x) \wedge z\{\mathsf{P}, \mathsf{PO}\}y \rightarrow \mathsf{CL}(z, x)]$; informally, a region z is close to a region x if another region y is *a priori* close to x and z is part of or partially overlaps y.

The subscript ap in the name of the relation $\mathsf{CL}_{\mathsf{ap}}(y, x)$ stands for "a priori". $\mathsf{CL}_{\mathsf{ap}}(y, x)$ is derived from background knowledge. In this paper we consider two sources of background knowledge, (1) a country's organization into different levels of administrative partitions (cf. section 4.2) and (2) tessellations of different granularity consisting of different types of functional regions (which may cross a country's borders).

Even though tessellations of different granularity may be organized as a system of partitions similar to that of administrative regions, this does not have to be the case. Our approach requires, however, that each administrative region must be related to

[7] This is consistent with Shariff, Egenhofer and Mark [8] who conclude that, for a large set of spatial-relation terms, topology is a more important parameter of the semantics than metric.

[8] Worboys [9] argues that for nearness the subject-referent dichotomy plays a dominant role in that the referent creates the scale in which the relation has context.

exactly one functional region. In the current implementation (cf. section 5.1) we use the weak notion of "located in" which is introduced as subsumed by "spatially related" – the most general RCC relation.

4.2 A Partially Ordered and Typed System of Partitions

Definition 1 uses the Boolean RCC function SUM and the RCC relation DR to reformulate the well-known notion of a partition in terms of RCC. The RCC function $\mathsf{SUM}_{i \in I}\, x_i$ is defined as $\forall z\, [\mathsf{C}(z, y) \leftrightarrow \mathsf{V}_{i \in I}\, \mathsf{C}(z, x_i)]$ for a region y [4]. As is customary, lower case letters are used for variables denoting individuals.

Definition 1 (Partition in RCC). A family of regions $(x_i)_{i \in I}$ is a partition of a region y if the following holds:

- $y = \mathsf{SUM}_{i \in I}\, x_i$ where I is a finite index set; this implies $\forall x_i\, \mathsf{P}(x_i, y)$;
- $\forall x_i \forall x_j\, \mathsf{DR}(x_i, x_j)$ for $i \neq j$;
- regions $(x_i)_{i \in I}$ are named for all $i \in I$.

We consider only a small subset of partitions, namely those whose elements are typed by kind of administrative region. For instance, $\mathsf{Community}(x_i)$ says that x_i is of type Community. Multiple typing of regions is not considered, that is, the concepts used for typing are mutually disjoint. Similarly, a given type is used for a single partition only. This allows distinguishing the partitions by their types.

In order to account for the different scales of social organization a *partial order* on the system of partitions in RCC is defined by comparing partitions with regard to their granularity.

Definition 2 (Partial Order on Typed Partitions). Let $C(x_i)_{i \in I}$ and $D(y_k)_{k \in K}$ be partitions of the same region of types C and D, respectively. We say that $C(x_i)_{i \in I}$ is *more fine-grained* than $D(y_k)_{k \in K}$, denoted by $C(x_i)_{i \in I} \preceq D(y_k)_{k \in K}$, if each element of $C(x_i)_{i \in I}$ is a (possibly improper) subset of an element of $D(y_k)_{k \in K}$. A partial order on typed partitions is reflexive, transitive and antisymmetric.

This means that each element of $D(y_k)_{k \in K}$ is partitioned by elements of $C(x_i)_{i \in I}$. For instance, $\mathsf{Community}(x_i)_{i \in I}$ and $\mathsf{District}(y_k)_{k \in K}$ are both typed partitions of a canton and each element of $\mathsf{District}(y_k)_{k \in K}$ is partitioned by elements of $\mathsf{Community}(x_i)_{i \in I}$.

Definition 3 (Minimal Partial Order on Typed Partitions). We say that a partial order on typed partitions is *minimal* with regard to a given conceptualization, denoted by $C(x_i)_{i \in I} \preceq_{\min} D(y_k)_{k \in K}$, if the conceptualization does not provide a type for any $(w_j)_{j \in J}$ such that $C(x_i)_{i \in I} \preceq (w_j)_{j \in J} \preceq D(y_k)_{k \in K}$. A minimal partial order on typed partitions is *intransitive*.

For instance, if a given conceptualization provides the administrative types District and Community, any partial order comprising a non-typed partition of intermediate

granularity is not minimal. Definition 3 excludes unwanted partitions such as those consisting of a mash of districts and communities. For further information cf. [3].

4.3 Refining the Formalism

The above introduced background knowledge can be used to formalize the notion of *a priori* closeness as shown in composition rule 2.

Composition rule 2. $\forall x_a \in (a_i)_{i \in I}\, \forall y_a \in (a_i)_{i \in I}\, \forall b \in (b_k)_{k \in K}\, \forall w\, [\mathsf{P}(x_a, b) \wedge y_a\{\mathsf{P}, \mathsf{EC}\}b \wedge \mathsf{LOC}(x_a, w) \wedge \mathsf{LOC}(y_a, w) \rightarrow \mathsf{CL}_{\mathsf{ap}}(y_a, x_a)]$; informally, a region y_a is *a priori* close to a region x_a, if (i) x_a and y_a belong to the same administrative partition $(a_i)_{i \in I}$ (e.g., both are communities); (ii) y_a is part of or borders the same region b of the next upper level of administrative partitions $(b_k)_{k \in K}$ (e.g., a district) of which x_a is part; and (iii) x_a and y_a are located (LOC) in the same functional region w of appropriate granularity; EC stands for "externally connected to".

Note that in composition rule 2 the scope of the quantifiers for x_a, y_a and b is limited to the elements of the respective partitions. This also applies to composition rule 1', a refinement of composition rule 1 which uses the consequence of composition rule 2 in the rule body. Composition rules 1' and 2 are implemented in our rule base.

Composition rule 1'. $\forall x_a \in (a_i)_{i \in I}\, \forall y_a \in (a_i)_{i \in I}\, \forall z\, [\mathsf{CL}_{\mathsf{ap}}(y_a, x_a) \wedge z\{\mathsf{P}, \mathsf{PO}\}y_a \rightarrow \mathsf{CL}(z, x_a)]$; informally, a region z is close to a region x_a of an administrative partition $(a_i)_{i \in I}$ if another region y_a of the same administrative partition is *a priori* close to x_a and z is part of or partially overlaps y_a.

5 Rewriting Queries That Use Local Expressions

5.1 Representing Topological Background Knowledge

For the evaluation of composition rules 1' and 2 three partitions of administrative units and a tessellation of functional micro regions are asserted as background knowledge in an OWL DL Knowledge Base ($\mathcal{KB}$), consisting of a TBox $\mathcal{T}$ and an ABox $\mathcal{A}$. Rules are implemented in a DL-safe SWRL rule base ($\mathcal{RB}$) (not shown). The notation used for the $\mathcal{KB}$ is adopted from [10]. The expressivity of the description language is $\mathcal{ALCHOIF}$. Note that the complexity of the approach is determined by DL complexity.

Partitions are represented in $\mathcal{T}$ by (anonymous) concepts that are made up of individual names, also called *nominals*, $\{a_1, \ldots, a_n\}$. Nominals are linked to types by axioms of the form $C \sqsubseteq \{a_1, \ldots, a_n\}$. In order to disallow multiple typing, the concepts used for typing are defined as mutually disjoint, i.e. $C \sqsubseteq \neg D$.

The subsumption hierarchy of RCC relations [4] is implemented as a hierarchy of roles. The role partOf and the roles subsumed by partOf are described as functional roles, thereby making sure that an individual region a_i can be part of a single region b_j only. This overrides the transitivity of the RCC relation $\mathsf{P}(x, y)$ and prevents, for instance, communities from being related to cantons (or to countries or continents if these were represented).

Disjunctions of RCC relations in the bodies of composition rules, such as {P, PO}, are represented by auxiliary roles subsuming the roles partOf and partiallyOverlaps, for instance. This has some similarity with the design of RCC-12 [11]. RCC-12 relations generalize the RCC-8 relations in such a way as to allow composition rules for being expressed as (non-disjunctive) Horn rules.

Partitions are asserted in $\mathcal{A}$ as $\mathsf{partOf}(a_i, b_j)$, or any of the roles subsumed by $\mathsf{partOf}(a_i, b_j)$, for all applicable $a_i \in \{a_1, \ldots, a_n\}$ and $b_j \in \{b_1, \ldots, b_m\}$. In so doing, $\mathcal{A}$ is closed with regard to nominals denoting administrative regions.[9] A minimal partial order on typed partitions (cf. section 4.2) is implemented by asserting $\mathsf{partOf}(a_i, b_j)$ or any of the roles subsumed by $\mathsf{partOf}(a_i, b_j)$ only for those pairs of individuals (a_i, b_j) for which holds $C(a_i)_{i \in I} \preceq_{\min} D(b_j)_{j \in J}$. All individuals in the ABox are asserted as being different from each other.

5.2 Rewriting Queries

Algorithm 1. Function **CLOSETO** computes $(Q \sqcap \exists \mathsf{closeTo}.\{a\})(z)$ from $\mathcal{KB}$ and $\mathcal{RB}$ using composition rules 1' and 2 (cf. section 4.3).

FUNCTION CLOSETO
INPUT: Knowledge Base $\mathcal{KB} = \{\mathcal{T}, \mathcal{A}\}$, Rule Base $\mathcal{RB}$, Concept Q, Individual a
OUTPUT: Set<Individual>

0. $U \leftarrow \emptyset$, $V \leftarrow \emptyset$, $W \leftarrow \emptyset$, $X \leftarrow \emptyset$, $Y \leftarrow \emptyset$, $Z \leftarrow \emptyset$
1. $\{b\} \leftarrow \{b \mid \mathcal{A} \models \mathsf{partOf}(a, b)\}$
2. $U \leftarrow \{u_{i \in I} \mid \mathcal{A} \models \mathsf{partOfOrExternallyConnectedTo}(u_i, b)\}$
3. $\{c\} \leftarrow \{c \mid \mathcal{A} \models \mathsf{locatedIn}(a, c)\}$
4. $V \leftarrow \{v_{j \in I} \mid \mathcal{A} \models \mathsf{locatedIn}(v_j, c)\}$
5. $Y \leftarrow U \cap V$
   FOR $(y_k \in Y;\ Y \neq \emptyset;\ Y \setminus y_k)$ {
6. $X \leftarrow X \cup \{x_{m \in M} \mid \mathcal{A} \models \mathsf{partOfOrPartiallyOverlaps}(x_m, y_k)\}$ }
7. $W \leftarrow \{w_{n \in N} \mid \mathcal{A} \models Q(w_n)\}$
8. $Z \leftarrow X \cap W$
9. **OUTPUT** Z

The terms used in a query reveal how a user conceptualizes a domain. Query concepts can, thus, be used to determine the scale on which spatial relations are to be

[9] Note that $\mathsf{partOf}(a_i, b_j)$ and the roles subsumed by $\mathsf{partOf}(a_i, b_j)$ are used for asserting partitions into administrative regions only.

evaluated. For the evaluation (cf. section 6), conjunctive queries of the form $\forall z\ [Q(z) \wedge \mathsf{CL}(z, a)]$ are used, which are expected to return the set of those individuals of type Q that are close to a given individual a. In this query, the type of individual a, for instance Community, determines the scale for the evaluation of $\mathsf{CL}(z, a)$.

Algorithm 1 show the steps (0–9) to take when rewriting a query. The query $\forall z\ [Q(z) \wedge \mathsf{CL}(z, a)]$ is implemented in DL by the concept assertion $(Q \sqcap \exists \mathsf{closeTo}.\{a\})(z)$. Given an ABox $\mathcal{A}$ and a concept description $Q \sqcap \exists \mathsf{closeTo}.\{a\}$, the retrieval problem is thus to find all individuals z in $\mathcal{A}$ such that $\mathcal{A} \models (Q \sqcap \exists \mathsf{closeTo}.\{a\})(z)$.

6 Evaluation

6.1 Material and Methods

We compare the results of two series of web searches using 170 pairs of conceptually (although not syntactically) consistent queries according to two different strategies. According to the first search strategy, the queries are entered into the search engine as a set of strings. According to the second search strategy, the queries are semantically rewritten and the resulting queries are fed into the search engine. The knowledge required to rewrite the queries is held in a consistent DL knowledge base and a DL-safe SWRL rule base as described in section 5.1. The knowledge base holds 12 concepts, 21 roles, 210 individuals, 603 concept assertions and 29,003 role assertions. Pellet 2.0[10] is used in order to rewrite the queries. Using Pellet 2.0 to reason on OWL DL knowledge bases returns sound and complete results [12]. Reasoning on SWRL rule bases is sound, but not necessarily complete [13]. However, the rewritten queries that were considered for our motivating example in section 3 were also complete. The search engine used for the comparison is GoForIt.[11]

In order to compare the searches, recall and precision are calculated. GoForIt is based on the Open Directory Project (ODP).[12] Different from ODP's search engine, however, GoForIt not only searches the directory's content, but also the categorized resources. This allows extracting all figures necessary for the calculation of recall and precision. The numbers of relevant resources in the result sets are found by summing up the figures in the relevant categories. To give an example, rewriting the query `<Gemeinden "in der Nähe von" Dietlikon>` (i.e. German for communities close to Dietlikon) returns the names of eight communities (cf. Fig. 1). The relevant categories of a search using the disjunction of these names are `Nürensdorf`, `Dübendorf`, `Rümlang`, `Wallisellen`, `Kloten`, `Wangen-Brüttisellen`, `Bassersdorf` and `Opfikon`. For the calculation of recall, the returned resources in these categories are related to the sum of all resources (not only of those found by the engine) of the same categories. We thus make the common

[10] http://clarkparsia.com/pellet
[11] http://www.goforit.com/
[12] http://www.dmoz.org/

assumption that manually categorized resources are more relevant than those found by a search algorithm. For the calculation of precision, the returned resources are related to the numbers of resources (whether relevant or not) in the result sets. A two-sided, pairwise t-test has been performed on the resulting recall values to show the significance of our results.

This analysis is complemented by measuring the time required to rewrite the queries.

6.2 Results

Searches without Query Rewriting. In this part of the evaluation the retrieval problem stated in section 3 is put in terms of the strings `<Gemeinden "in der Nähe von" Dietlikon>`. Similar queries are framed for the remaining 169 communities in the canton of Zurich.[13] The results from web searches using such strings are summarized in Table 1. They are discussed below.

Table 1. Results from searches without query rewriting (n = 170)

	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

Searches with Query Rewriting. In this part of the evaluation the retrieval problem is put in terms of the following SPARQL query [14]:

```
SELECT ?z
WHERE {
  ?z rdf:type exp:Community .
  ?z rdf:type [a owl:Restriction;
  owl:onProperty exp:closeTo;
  owl:hasValue exp:Dietlikon] .
}
```

The result of the SPARQL query is fed into the search engine: `<Nürensdorf OR Dübendorf OR Rümlang OR Wallisellen OR Kloten OR Wangen-Brüttisellen OR Bassersdorf>`. Similar queries are framed for the remaining 169 communities. The numbers of community names resulting from query rewriting range between 6 and 24. The results from the searches are summarized in Table 2.

[13] Note that we excluded the community of Zurich from the analysis. The rewriting algorithm returns intuitively satisfactory results for 170 communities, but not for Zurich. It seems that for communities like Zurich a topological model also has to take into account the impact of urban agglomeration.

Table 2. Results from searches with query rewriting (n = 170)

	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

6.3 Discussion

Recall of all searches without query rewriting is low. Only 17 out of 170 searches return a relevant match. The reason for this is that GoForIt does not return resources of entities that are "close to" the reference places as does Google for government agencies and commercial enterprises in urban areas (cf. section 1). Precision is undefined for 102 searches which makes the calculation of a meaningful average infeasible.

Query rewriting significantly ($p < 0.01$) increases recall of the searches. Precision is defined for all searches with query rewriting, at a consistently low level, however. When appraising precision one should keep in mind that the method of calculation disregards the ranking algorithm of the search engine. Precision of the n-best results is much higher. All 170 searches are located in the quadrant of the recall $\times$ precision matrix (not shown) that is far from the precision axis (i.e. recall > 0.5) and close to the recall axis (i.e. precision < 0.5). According to Salton and McGill [15], this characterizes broad searches put in general terms.

Overall response time is determined by the time required to rewrite the queries. When using an X86-based PC with a clock rate of 2,533 MHz and a Random Access Memory of 4 GB to operate on the knowledge/rule base described in section 5.1, the time required for query rewriting ranges between 5,608 ms and 19,452 ms (6,317.54 ms on the average). This is acceptable except for six queries which take over 10,000 ms to be rewritten.

The evaluation assumes that query rewriting properly interprets the intended meaning of the expression "close to" in the given context (which remains to be seen). However, even if our approach approximated the meaning of "close to" only roughly, it would still be useful to improve the searches. This can be seen from a comparison of the average total matches in Tables 1 and 2.

7 Related Work

7.1 Administrative Boundaries Influence the Perception of Distance

Maki [16] showed that the affiliation to a category, such as a state, plays an important role in human perception of locations. Subjects should decide about the location of two cities regarding their orientation east-west. If the cities in question belong to different states, the reaction times were significantly shorter than with cities which

belong to the same state. The term "categorization effect" refers to the fact that human beings are able to judge faster about entities on a continuum if they can make use of category information.

Carbon and Leder [17] showed that the membership to different political systems, structures or hierarchies influences the estimation of distance between two cities. In their experimental setting, subjects should estimate distances between cities east and west of the former border inside of Germany. Compared to pairs inside the same part of the former republic, distances were overestimated if the cities in question belonged to different parts.

Based on investigations in natural-language corpora, Hois and Kutz [18] are providing parameters which influence the human perception of space. Among these is "domain-specific knowledge of entities" which refers to things such as granularity. Granularity in our approach is modeled via different layers of administrative regions.

7.2 Using Local Expressions in Web Searches

Mark and Egenhofer [19] describe an experiment to test how people think about spatial relations between unbranched lines and simply connected regions. For the predicates "the road crosses the park" and "the road goes into the park" there was a great deal of consensus among the subjects. The authors conclude that the so-called 9-intersection model forms a sound basis for characterizing line-region relations and that many spatial relations can be well-represented by particular subsets of the primitives differentiated by the 9-intersection model.

Different from the approach described here, Mark and Egenhofer [19] use verbs to term natural language predicates and not prepositional phrases. This is reasonable, because in their cases, verbs catch the intuition of the predicates better than any other word class. Independent of the word class used, their results suggest that natural language predicates can, in principle, be aligned with spatial relations as identified by a mathematical model. This supports a similar suggestion for spatial relations between simply connected regions made by the approach described here.

The European SPIRIT project addressed the shortcomings of web search facilities when considering geographical context [20]. It developed methods supporting spatially-aware information retrieval on the Internet. The core component of the system is a geographical ontology that provides a model of the terminology and structure of geographic space. The geographical ontology supports "part-of", "contains", "overlap" and "adjacency" relations between geographic places. Together with the disambiguated place name such relations are used to derive the desired geographical search extent for the query. While "part-of", "contains", "overlap" and "adjacency" can be mapped onto the RCC relations, they are arbitrarily chosen and do not form a jointly exhaustive and pairwise disjoint set of relations. Relations that do not fall into any of the four categories (e.g., "disconnected from") and relations that extend RCC (e.g., "close to") are undefined.

Bishr [21] proposes to encode spatial inferences in the Semantic Web Rule Language (SWRL) [22]. Even though not explicitly mentioned, the examples are provided in an RCC-like style. The proposal can, in principle, be aligned with the

approach presented here. Different from [21], however, we introduce additional relations and provide an implementation.

Schokaert, De Cock and Kerre [23] (in [24]) suggest augmenting the structured information available to a local search service, such as Google Maps, with information extracted from the web. They show how nearness information in natural language and information about the surrounding neighborhood of a place can be translated into fuzzy restrictions and how such fuzzy restrictions can be used to estimate the location of a place with an unknown address.

While the idea of augmenting the structured information available to a local search service with information extracted from semi- and unstructured data, i.e. documents on the web, is appealing, it requires that the latter is available in abundance. The "vast amount" [23] of data addressed by the authors, together with the kinds of examples they provide, suggest that their approach is targeted on mass searches. In our case, the resources on the web, which could possibly be used to augment the searches, are scarce (cf. section 6).

8 Conclusion and Outlook

We introduced an approach to rewrite queries for web searches that use local expressions. Query rewriting makes use of topological background knowledge that is implemented in an OWL DL knowledge base and a DL-safe SWRL rule base. Applying the approach to searches for communities which are "close to" a reference place shows that query rewriting significantly improves recall of the searches.

The spatial relations between two simply connected regions identified by the 9-intersection model mentioned in section 7 equal the RCC-8 relations. To the best of our knowledge, experiments testing natural language predicates for compliance with these relations in a way similar to that described for unbranched lines and simply connected regions [19] have not been performed so far. Likewise, no experiments have been performed with the newly introduced relation "close to". Whether the described approach is empirically well founded or not remains to be seen.

Our approach requires that topologies of administrative units are available in RCC. State-of-the-art geographic information systems (GIS) and spatial databases provide ways and means to compute such topologies from GIS layers. In Switzerland the relevant GIS layers can be downloaded from the website of the Swiss Federal Statistical Office.[14] Other European countries such as the United Kingdom and Germany offer similar services. Technically, the approach is, thus, applicable to many countries. Whether the semantics of the relation "close to", expressed as rules applied to the generated topologies, differs between countries remains to be seen.

The current prototype operates on topological knowledge that covers a part of Switzerland. Since the knowledge base grows by the square of the number of regions asserted, we expect the performance to decrease when extending the coverage area. This applies even though an off the shelf PC was used for the evaluation, which could

[14] http://www.bfs.admin.ch/bfs/portal/de/index/dienstleistungen/geostat/datenbeschreibung/generalisierte_gemeindegrenzen.html

easily be replaced by a faster one. Future work will explore ways of dealing with this expected decrease of performance. This will include distribution of knowledge bases and outsourcing of individuals in a database or a triple store which are known to scale better than in-memory storage structures. An even better way might be to move expensive knowledge processing from run-time to design-time. This requires that search engines are enabled to use topological background knowledge when crawling the web and indexing resources. Operating on index entries such as `<Nürensdorf: "close to" Dietlikon>` at run-time is expected to be much faster than rewriting queries.

The approach presented in this paper distinguishes between the basic formalism and the way how background knowledge is used in order to ground the relation $\mathsf{CL}_{\mathsf{ap}}(y, x)$. This separation clears the way for using alternate sources of background knowledge. Put the other way round, it facilitates the use of alternate approaches to compute proximity on the basis of the background knowledge provided in this work. Accordingly, we intend to use travel time as calculated by a route planning algorithm to estimate spatial closeness and to relate the results to those obtained by the approach presented here in the near future.

Acknowledgements. This research was funded by and conducted in cooperation with the Swiss Federal Office for the Environment (FOEN) and the Swiss National Science Foundation (SNF).

References

1. Sanderson, M., Kohler, J.: Analyzing geographic queries. In: Proceedings of the Workshop on Geographic Information Retrieval (SIGIR). ACM Press, New York (2004)
2. Wang, L., Wang, C., Xie, X., Forman, J., Lu, Y., Ma, W.-Y., Li, Y.: Detecting Dominant Locations from Search Queries. In: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR'05). ACM, New York, NY, USA (2005)
3. Randell, D.A., Cui, Z., Cohn, A.G.: A Spatial Logic based on Regions and Connections. In: Nebel, B., Rich, C., Swartout, W. (eds.) Proceedings of the 3rd International Conference on Principles of Knowledge Representation and Reasoning (KR'92), pp. 165–176. Morgan Kaufmann, San Mateo, CA (1992)
4. Motik, B., Sattler, U., Studer, R.: Query Answering for OWL-DL with Rules. J. Web Semant. 549–563 (2004)
5. Motik, B., Horrocks, I., Rosati, R. Sattler, U.: Can OWL and Logic Programming Live Together Happily Ever After? In: Cruz, I.F. et al. (eds.) Proceedings of the 5th International Semantic Web Conference (ISWC 2006). LNCS, vol. 4273, pp. 501–514. Springer, Heidelberg (2006)
6. Bundesamt für Statistik: Räumliche Gliederungen der Schweiz, http://www.bfs.admin.ch/bfs/portal/de/index/regionen/11/geo.html (2011)
7. Shariff, A.R., Egenhofer, M., Mark, D.: Natural-Language Spatial Relations Between Linear and Areal Objects: The Topology and Metric of English-Language Terms. International Journal of Geographical Information Science 12 (3), 215–246 (1998)
8. Worboys, M.F.: Nearness Relations in Environmental Space, International Journal of Geographical Information Science 15 (7), 633–651 (2001)

9. Grütter, R., Scharrenbach, T., Waldvogel, B.: Vague Spatio-Thematic Query Processing – A Qualitative Approach to Spatial Closeness. Transactions in GIS 14 (2), 97–109 (2010)
10. Baader, F., Nutt, W.: Basic Description Logics. In: Baader, F., Calvanese, D., McGuinness, D.L., Nardi, D., Patel-Schneider, P.F. (eds.) The Description Logic Handbook: Theory, Implementation and Applications (2nd ed.), pp. 47–104. Cambridge University Press (2007)
11. Schockaert, S.: Reasoning About Fuzzy Temporal and Spatial Information From the Web. Ph.D. thesis, Ghent University (2008)
12. Sirin, E., Parsia, B., Grau, B.C., Kalyanpur, A., Katz, Y.: Pellet: A Practical OWL-DL Reasoner. J. Web Semant. 51-53 (2007)
13. Parsia, B.: Understanding SWRL (Part 3): Some tricky bits. Weblog Clark & Parsia, LLC, Thursday, September 13th, 2007, http://weblog.clarkparsia.com/2007/09/13/understanding-swrl-part-3-some-tricky-bits/
14. Prud'hommeaux, E., Seaborne, A.: SPARQL Query Language for RDF. W3C Recommendation 15 January 2008, http://www.w3.org/TR/2008/REC-rdf-sparql-query-20080115/
15. Salton, G., McGill, M.J.: Introduction to modern information retrieval. McGraw-Hill, New York, NY (1983)
16. Maki, R.H.: Categorization and Distance Effects With Spatial Linear Orders. Journal of Experimental Psychology: Human Learning and Memory 7 (1), 15–32 (1981)
17. Carbon, C.-C., Leder, H.: The Wall Inside the Brain: Overestimation of Distances Crossing the Former Iron Curtain. Psychonomic Bulletin & Review 12 (4), 746–750 (2005)
18. Hois, J., Kutz, O.: Natural Language Meets Spatial Calculi. In: Freksa, C., Newcombe, N.S., Gärdenfors, P., Wölfl, S. (eds.) Spatial Cognition VI. Learning, Reasoning, and Talking about Space. LNCS, vol. 5248, pp. 266–282. Springer, Heidelberg (2008)
19. Mark, D.M., Egenhofer, M.J.: Modeling Spatial Relations Between Lines and Regions: Combining Formal Mathematical Models and Human Subjects Testing. In: Egenhofer, M.J., Mark, D.M., Herring, J. (eds.) The 9-Intersection: Formalism and its Use for Natural-Language Spatial Predicates. Technical Report 94-1, National Center for Geographic Information and Analysis, University of California, Santa Barbara, CA (1994)
20. Jones, C.B., Purves, R., Ruas, A., Sanderson, M., Sester, M., van Kreveld, M., Weibel, R.: Spatial Information Retrieval and Geographical Ontologies: An Overview of the SPIRIT Project. In: Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 387-388. ACM Press (2002)
21. Bishr, Y.: Geospatial Semantic Web. In: Rana, S., Sharma, J. (eds.) Frontiers of Geographic Information Technology, pp. 139–154. Springer, Berlin Heidelberg (2006)
22. Horrocks, I., Patel-Schneider, P.F., Boley, H., Tabet, S., Grosof, B., Dean, M.: SWRL: A Semantic Web Rule Language Combining OWL and RuleML. W3C Member Submission 21 May 2004. World Wide Web Consortium, http://www.w3.org/Submission/SWRL/
23. Schockaert, S., De Cock, M., Kerre, E.E.: Location approximation for local search services using natural language hints. International Journal of Geographic Information Science 22 (3), 315–336 (2008)
24. Jones, C.B., Purves, R.S.: Special Issue: Geographical Information Retrieval. International Journal of Geographic Information Science 22 (3), 219–360 (2008)