Reasoning about distances

Frank Wolter and **Michael Zakharyaschev**

Department of Computer Science, University of Liverpool,
Liverpool L69 7ZF, U.K.;
Department of Computer Science, King's College,
Strand, London WC2R 2LS, U.K.
(emails: frank@csc.liv.ac.uk, mz@dcs.kcl.ac.uk)

Abstract

The paper presents a novel expressive logic-based formalism intended for reasoning about numerical distances. We investigate its computational properties (in particular, show that it is EXPTIME-complete) and devise a tableau-based satisfiability-checking algorithm. To be able to express knowledge about implicit or unknown distances, we then extend the language with variables ranging over distances and prove that the resulting logic is decidable as well.

1 Introduction

The numerical notion of *distance* or *metric* as a measure of closeness or similarity (in a great variety of senses) plays a fundamental role in many branches of science—including, naturally, knowledge representation, artificial intelligence and informatics in general. Recent examples are such different fields as spatial KR&R and geographic information systems [Cohn and Hazarika, 2001; Adam and Gangopadhyay, 1997], computational molecular biology [Setubal and Meidanis, 1997; Clote and Backofen, 2000], text processing [Salton, 1989] and data mining [Dunham, 2003]. Unlike classical physics and the corresponding mathematical models, the 'distance spaces' arising in these applications may be quite different from the standard Euclidean spaces, e.g., they are often finite and do not necessarily satisfy all of the metric axioms. Another important feature is that properties of *only some* objects in the space may be known, and the task is to use the available information to determine properties of other elements of interest. Traditional methods of dealing with metric data are mostly numerical or 'sub-symbolic' (e.g., graph-theoretic or probabilistic) and usually do not involve any kind of (automated) reasoning which is based upon symbolic, explicit knowledge representation. *The main aim of this paper is to present a new expressive logic-based formalism intended for reasoning about distances and to provide it with a tableau decision algorithm.*

The language we propose is—as far as its algorithmic properties and expressiveness are concerned—a compromise between full first-order and propositional languages: it contains object names and unary atomic predicates from which complex predicates are constructed by means of the Boolean operators and restricted quantifiers like, e.g., $\forall^{<r}$ ('for all objects within a circle of radius r') and $\exists^{<r}$ ('for some object within a circle of radius r'). Thus, $\exists^{<r}\mathsf{P}$ denotes the set of all objects each of which is located at distance $< r$ from at least one object with property P. To illustrate the expressive power of the language and to convince the reader that it can be useful, we give three concrete examples.

Example 1. Consider a space of proteins (i.e., sequences of amino acids); some of them are well studied, but properties of the majority are not known. Using one of the available techniques, we can define a 'similarity measure' between proteins which gives rise to a certain 'distance' (perhaps, metric) space. One of the typical problems is to recognise whether a given protein p has some property P, e.g., whether it contains an EF-hand motif. The usual approach of comparing the structure of p with that of 'closely located' proteins whose properties are known can be supplemented with a knowledge base composed by an expert and containing information like

- $\exists^{\leq 5}\mathsf{Q} \sqcap \mathsf{P} \doteq \bot$ — 'no protein located at distance ≤ 5 from a protein with a property Q has P,'
- $\exists^{\leq 2}\{\ell\} \sqsubseteq \mathsf{R}$ — 'all proteins located at distance ≤ 2 from a particular protein ℓ have a property R,' etc.

Note that, in view of incomplete information about properties of elements of the distance space, to answer the query whether p can have P, we have to run a satisfiability-checking algorithm rather than a model-checking one.

Example 2. Suppose now that we are trying to organise a web site for selling books (or any other goods) on-line. One way of classifying the existing books is to introduce a measure of closeness of the contents of *some* of them. We then obtain a database containing items of the form $\boldsymbol{\delta}(b_1, b_2) = a$, if the distance between books b_1 and b_2 is a, and, say, $\boldsymbol{\delta}(b_1, b_2) > c$, for some number c, if the distance between b_1 and b_2 is only known to be greater than c. Now we can built a sort of terminology by saying:

- $\mathit{Harry\_Potter} \sqsubseteq \mathsf{Fiction}$,
- $\mathsf{Fiction} \sqcap \exists^{\leq 3}\mathsf{Science} \sqsubseteq \mathsf{Science\_Fiction}$,
- $\mathsf{Science\_Fiction} \sqsubseteq \forall^{\leq 2}\neg\mathsf{Science}$,
- $\exists^{<1}\mathsf{AI} \sqcap \exists^{\leq 1}\mathsf{Science\_Fiction} \not\doteq \bot$,
- $\forall^{\leq 1}\mathsf{Computer\_Science} \sqsubseteq \mathsf{Science}$, etc.,

and deduce, for instance, that AI is not subsumed by Computer Science.

Example 3. Perhaps the most natural application of this kind of formalism is spatial representation and reasoning. We can speak in terms of 'absolute' distances:

- $\delta(London, Leipzig) > 100$,
- $house \sqsubseteq \exists^{<1}\forall^{\leq 0.5}(\mathsf{shop} \sqcup \mathsf{restaurant})$ — 'the house is located within 1 mile distance from a shopping centre of radius 0.5,'
- $\mathsf{doughnut} \doteq \exists^{\leq 10}\{centre\} \sqcap \neg\exists^{<6}\{centre\}$.

We can also deal with distances given implicitly by saying, for example,

- $\delta(London, Leipzig) > 3 \cdot \delta(Liverpool, Manchester)$,
- $my\_house \sqsubseteq \exists^{<x}\mathsf{college}$, $\delta(Dov's\_house, \mathsf{college}) \doteq x$ ('I do not want to have a house that is located farther from my college than Dov's house')

without referring to concrete distances between the cities and perhaps having no idea where Dov's house is.

The approach to reasoning about distances we advocate in this paper has grown up from four main sources. First, extensive research has been done by the (spatial/constraint) database community [Revesz, 2002] aiming to deal efficiently with large sets of data which come equipped with some kind of distance (or similarity) measure; consult, e.g., the literature on 'similarity search' and 'k-nearest neighbour queries' [Bozkaya and Özsoyoglu, 1999; Hjaltason and Samet, 1999]. Although this paper does not contribute to this field,[1] we were driven by the idea that it may be useful to extend the 'evaluation-based' methods by automated reasoning services. As was mentioned above, complete information about models based on distance spaces, and even distances themselves, are often not available, which means that the quality of solutions provided by the database technology can be significantly improved by using explicit knowledge of the field (or even conjectures) supported by appropriate reasoning procedures.

Two approaches to extend the database technology by means of KR&R have been developed by the qualitative spatial reasoning community. One is based upon the representation of spatial data by means of constraint systems [Renz and Nebel, 1999], while the other uses first-order theories, e.g., [Randell *et al.*, 1992]. Constraint systems are usually very efficient (at most NP-complete); however, their expressive power is rather limited (for instance, they do not allow any kind of quantification). On the other hand, the unrestricted quantification of first-order representations is a serious obstacle for their use in practical reasoning systems.

In a sense, our approach is similar to that taken by the description logic community which represents conceptual knowledge by means of restricted quantification (a kind of modal necessity and possibility operators) and uses reasonably effective tableau-based procedures even for expressive and complex (say, EXPTIME-complete) formalisms [Horrocks, 1998]. The chief difference from description logics is that we have to integrate rather sophisticated reasoning about *quantitative data* into a logic-based framework.

Logics of distance spaces were introduced in [Sturm *et al.*, 2000; Kutz *et al.*, 2003], where our main concern was to find out whether such logics can be decidable. Unfortunately, the logics constructed in these papers were not supported by effective reasoning machinery. The aim of this paper is more practical. To begin with, in Section 2, we present an application oriented variant $\mathcal{MS}^{\lessgtr}_{\leq}[\mathbb{Q}^+]$ of a distance logic (containing location constants, set variables and restricted quantifiers) and show that the satisfiability problem for its formulas in metric (and some weaker distance) spaces is EXPTIME-complete. As distance spaces used in computer science and AI are often finite (e.g., data with a similarity measure, a model of a railway system, etc.), we consider also the problem of satisfiability of $\mathcal{MS}^{\lessgtr}_{\leq}[\mathbb{Q}^+]$ knowledge bases in finite spaces. We show that the language does not 'feel' the difference between finite and infinite spaces: all the results presented for satisfiability in arbitrary spaces hold for finite spaces as well. (Unfortunately, if we restrict attention to the two-dimensional Euclidean space or its subspaces, then the satisfiability problem becomes undecidable.) Then, in Section 3, we devise a tableau-based satisfiability checking algorithm for this logic. An interesting and quite unexpected consequence of the proofs of these results is the following observation. The full $\mathcal{MS}^{\lessgtr}_{\leq}[\mathbb{Q}^+]$ contains both operators $\exists^{\leq a}$ and $\exists^{<a}$ which together allow assertions like

- $house \sqsubseteq \exists^{\leq 1}\mathsf{shop} \sqcap \neg\exists^{<1}\mathsf{shop}$ — 'the closest shop is located in precisely 1 mile distance from the house.'

But in everyday life we often do not pay attention to the boundaries of spatial regions or treat them in a sort of vague manner. What happens if we keep only one of the operators $\exists^{\leq a}$ or $\exists^{<a}$ in our language? Do the two resulting fragments behave differently, or to put it another way, does our language feel the difference between regions *with* and *without* boundaries? To our surprise, the answer turns out to be 'NO' (for a precise formulation of the result see Theorem 7).

To be able to express knowledge about implicit or unknown distances, in Section 4 we extend $\mathcal{MS}^{\lessgtr}_{\leq}[\mathbb{Q}^+]$ with variables ranging over distances and prove that the satisfiability problem for the resulting language is decidable.

Because of lack of space we had to omit all (rather nontrivial and lengthy) proofs. They are available at `http://www.csc.liv.ac.uk/~frank`). Here we only say that the complexity result as well as the proof of the correctness and completeness of the tableau algorithm are based on a representation of metric spaces in the form of relational structures *à la* Kripke frames. This representation allows us to use advanced techniques from modal and description logics.

2 The logic $\mathcal{MS}^{\lessgtr}_{\leq}[M]$

Let M be a non-empty subset of positive rational numbers (i.e., a subset of $\mathbb{Q}^+$). Denote by $\mathcal{MS}^{\lessgtr}_{\leq}[M]$ the language with the following alphabet and term and formula formation rules. The alphabet consists of

- an infinite list of *set* (or *region*) *variables* $X_1, X_2, \ldots$;
- an infinite list of *location constants* $\ell_1, \ell_2 \ldots$,

[1] Note, however, that our logics can express some of the clustering conditions.

- the *set constant* (or *nominal*) $\{\ell_i\}$, for every ℓ_i,
- binary *distance* ($\boldsymbol{\delta}$), *equality* ($\doteq$) and *membership* ($\sqsubseteq$) *predicates*;
- the *Boolean operators* $\sqcap$, $\neg$ (and their derivatives $\sqcup$, $\top$ and $\bot$);
- two *distance quantifiers* $\exists^{<a}$, $\exists^{\leq a}$ and their duals $\forall^{<a}$, $\forall^{\leq a}$, for every $a \in M$;
- two *universal quantifiers* $\exists$ and $\forall$.

Set terms s of $\mathcal{MS}^{\lessgtr}_{\leq}[M]$ are defined as

$$s ::= X_i \mid \{\ell_i\} \mid \top \mid \bot \mid \neg s \mid s_1 \sqcap s_2 \mid s_1 \sqcup s_2 \mid \exists^{<a} s \mid \exists^{\leq a} s \mid \forall^{<a} s \mid \forall^{\leq a} s \mid \exists s \mid \forall s,$$

and the *formulas* of $\mathcal{MS}^{\lessgtr}_{\leq}[M]$ are

- $\ell \sqsubseteq s$, where ℓ is a location constant and s a set term,
- $s \doteq t$ and $s \neq t$, where s and t are set terms,
- and all '*distance atoms*' of the form

  $$\boldsymbol{\delta}(k,\ell) < a, \;\; \boldsymbol{\delta}(k,\ell) \leq a, \;\; \boldsymbol{\delta}(k,\ell) > a, \;\; \boldsymbol{\delta}(k,\ell) \geq a,$$

  where k, ℓ are location constants and $a \in M$.

We use $s_1 \sqsubseteq s_2$ as an abbreviation for $s_1 \sqcap s_2 \doteq s_1$.

The intended semantics for this language is defined as follows. An $\mathcal{MS}^{\lessgtr}_{\leq}[M]$-*model* is a structure of the form

$$\mathfrak{B} = \left\langle W, d, X_1^{\mathfrak{B}}, X_2^{\mathfrak{B}}, \ldots, \ell_1^{\mathfrak{B}}, \ell_2^{\mathfrak{B}} \ldots \right\rangle,$$

where $\langle W, d\rangle$ is a *metric space* with a *distance function* d satisfying the axioms

$$d(x,y) = 0 \text{ iff } x = y, \tag{1}$$
$$d(x,z) \leq d(x,y) + d(y,z), \tag{2}$$
$$d(x,y) = d(y,x), \tag{3}$$

for all $x, y, z \in W$, $X_i^{\mathfrak{B}} \subseteq W$, and $\ell_i^{\mathfrak{B}} \in W$. Thus, $\mathfrak{B}$ defines explicitly the values of set variables and location constants. (Although we do not make the unique name assumption—i.e., $\ell_i^{\mathfrak{B}} = \ell_j^{\mathfrak{B}}$, $i \neq j$, is not prohibited—it is easy to force different location constants to denote different points by including $\{\ell_i\} \sqsubseteq \neg\{\ell_j\}$ into the knowledge base.) The *values* $s^{\mathfrak{B}}$ of other $\mathcal{MS}^{\lessgtr}_{\leq}[M]$-terms s in $\mathfrak{B}$ are computed inductively as follows:

$$\top^{\mathfrak{B}} = W, \qquad \bot^{\mathfrak{B}} = \emptyset,$$
$$\{\ell_i\}^{\mathfrak{B}} = \{\ell_i^{\mathfrak{B}}\}, \qquad (s_1 \sqcap s_2)^{\mathfrak{B}} = s_1^{\mathfrak{B}} \cap s_2^{\mathfrak{B}},$$
$$(s_1 \sqcup s_2)^{\mathfrak{B}} = s_1^{\mathfrak{B}} \cup s_2^{\mathfrak{B}}, \qquad (\neg s)^{\mathfrak{B}} = W - s^{\mathfrak{B}},$$

$$(\exists^{\leq a} s)^{\mathfrak{B}} = \{x \in W : \exists y \in W \left(d(x,y) \leq a \wedge y \in s^{\mathfrak{B}}\right)\},$$
$$(\exists^{< a} s)^{\mathfrak{B}} = \{x \in W : \exists y \in W \left(d(x,y) < a \wedge y \in s^{\mathfrak{B}}\right)\},$$
$$(\forall^{\leq a} s)^{\mathfrak{B}} = \{x \in W : \forall y \in W \left(d(x,y) \leq a \rightarrow y \in s^{\mathfrak{B}}\right)\},$$
$$(\forall^{< a} s)^{\mathfrak{B}} = \{x \in W : \forall y \in W \left(d(x,y) < a \rightarrow y \in s^{\mathfrak{B}}\right)\},$$
$$(\exists s)^{\mathfrak{B}} = \{x \in W : \exists y \in s^{\mathfrak{B}}\},$$
$$(\forall s)^{\mathfrak{B}} = \{x \in W : \forall y \in W \; y \in s^{\mathfrak{B}}\}.$$

The *truth-relation* $\mathfrak{B} \models \varphi$, where φ is an $\mathcal{MS}^{\lessgtr}_{\leq}[M]$-formula, is defined in the natural way:

- $\mathfrak{B} \models \ell \sqsubseteq s$ iff $\ell^{\mathfrak{B}} \in s^{\mathfrak{B}}$,
- $\mathfrak{B} \models s_1 \doteq s_2$ iff $s_1^{\mathfrak{B}} = s_2^{\mathfrak{B}}$ (and $\mathfrak{B} \models s_1 \neq s_2$ iff $s_1^{\mathfrak{B}} \neq s_2^{\mathfrak{B}}$),
- $\mathfrak{B} \models \boldsymbol{\delta}(k,\ell) \leq a$ iff $d(k^{\mathfrak{B}}, \ell^{\mathfrak{B}}) \leq a$,
- $\mathfrak{B} \models \boldsymbol{\delta}(k,\ell) < a$ iff $d(k^{\mathfrak{B}}, \ell^{\mathfrak{B}}) < a$, and similar for $\geq$ and $>$.

An $\mathcal{MS}^{\lessgtr}_{\leq}[M]$-*knowledge base* is a finite set of $\mathcal{MS}^{\lessgtr}_{\leq}[M]$-formulas. A knowledge base Σ is called *consistent* (or *satisfiable*) if there exists an $\mathcal{MS}^{\lessgtr}_{\leq}[M]$-model $\mathfrak{B}$ such that $\mathfrak{B} \models \varphi$ for all $\varphi \in \Sigma$; in this case we write $\mathfrak{B} \models \Sigma$.

The language $\mathcal{MS}^{\lessgtr}_{\leq}[M]$ is a proper fragment of the language $\mathcal{MS}[M]$ introduced in [Kutz *et al.*, 2003]. While $\mathcal{MS}[M]$ has the interesting property of being of the same expressive power as the two-variable fragment of first-order logic over metric spaces, reasoning in $\mathcal{MS}[M]$ is undecidable. [Kutz *et al.*, 2003] investigates in-depth the fragment $\mathcal{MS}^{\sharp}[M]$ of $\mathcal{MS}[M]$ with distance operators $\exists^{\leq a}$, $\exists^{>a}$ ('for some object outside a circle of radius a') and their duals. In particular, various decidability results are obtained. The expressivity of $\mathcal{MS}^{\lessgtr}_{\leq}[M]$ and $\mathcal{MS}^{\sharp}[M]$ is incomparable and the mathematics required to analyse these logics differs considerably. The tableau-algorithm presented here shows that $\mathcal{MS}^{\lessgtr}_{\leq}[M]$ is probably more suitable for efficient reasoning than $\mathcal{MS}^{\sharp}[M]$.

Our first result establishes the decidability and the computational complexity of the satisfiability problem for $\mathcal{MS}^{\lessgtr}_{\leq}[M]$-knowledge bases in metric spaces:

Theorem 4. *The satisfiability problem for $\mathcal{MS}^{\lessgtr}_{\leq}[M]$-knowledge bases in arbitrary metric spaces is EXPTIME-complete (even if the parameters from M are coded in binary).*

Actually the same result can be proved for satisfiability in weaker distance spaces, say, the class of spaces which do not necessarily satisfy the triangular inequality (2) or the symmetry axiom (3). The proof of this and some other theorems in this paper are based on a rather sophisticated representation of distance spaces in the form of relational structures *à la* Kripke frames. The EXPTIME satisfiability-checking algorithm is basically an elimination procedure which iteratively deletes from the set of all 1-types over a sufficiently small set of terms all those that contain distance quantifiers without witnesses.

According to the next result, the logic of metric spaces introduced above has the finite model property, and so Theorem 4 holds for the class of finite metric spaces as well (the third equivalence actually follows immediately from the fact that any finite metric space is isometric to a subspace of some L_∞^n, $n \in \mathbb{N}^+$ [Matousek, 2002]).

Theorem 5. *The following conditions are equivalent for any $\mathcal{MS}^{\lessgtr}_{\leq}[M]$-knowledge base Σ:*

- *Σ is satisfiable in a metric space;*
- *Σ is satisfiable in a finite metric space;*
- *Σ is satisfiable in a finite subspace of some space $L_\infty^n =$*

$(\mathbb{R}^n, d^n_\infty)$, *where*

$$d^n_\infty((a_1,\ldots,a_n),(b_1,\ldots,b_n)) = \max\{|a_i - b_i| : 1 \le i \le n\}.$$

It turns out, however, that these positive results do not hold if the intended metric space is the two-dimensional Euclidean space or the class of its subspaces (this result is proved by a reduction of the undecidable $\mathbb{N} \times \mathbb{N}$-tiling problem):

Theorem 6. (i) *Satisfiability of $\mathcal{MS}^{\lessgtr}_{\leq}[\{1,2\}]$-knowledge bases in subspaces of $\mathbb{R} \times \mathbb{R}$ is undecidable.*

(ii) *The satisfiability problem for $\mathcal{MS}^{\lessgtr}_{\leq}[\{1,2\}]$-knowledge bases in $\mathbb{R} \times \mathbb{R}$ is undecidable.*

Denote by $\mathcal{MS}^{<}[\mathbb{Q}^+]$ ($\mathcal{MS}^{\leq}[\mathbb{Q}^+]$) the fragment of $\mathcal{MS}^{\lessgtr}_{\leq}[\mathbb{Q}^+]$ without the quantifiers $\exists^{\leq a}$ and $\forall^{\leq a}$ ($\exists^{<a}$ and $\forall^{<a}$, respectively).

Theorem 7. *Suppose that Σ is a knowledge base over $\mathcal{MS}^{\leq}[\mathbb{Q}^+]$ and Σ' results from Σ by replacing all occurrences of $\forall^{\leq a}$ with $\forall^{<a}$ and $\exists^{\leq a}$ with $\exists^{<a}$. Then Σ is satisfiable iff Σ' is satisfiable.*

3 The tableau algorithm

Our tableau algorithm is closely related to tableau-based decision procedures from the field of description logic [Baader and Sattler, 2000]. However, in addition to the standard methodology of description logic, we have to take care of the 'arithmetical meaning' of the quantifiers.

Suppose that a knowledge base Σ is given. To simplify presentation, we will assume that Σ is formulated in the language $\mathcal{MS}^{\lessgtr}_{\leq}[\mathbb{N}^+]$ *without universal quantifiers* $\forall$ and $\exists$. (Knowledge bases containing rational numbers can be transformed into knowledge bases over $\mathbb{N}$ without effecting their satisfiability.)

Denote by $trm(\Sigma)$ the set of set terms occurring in Σ. Without loss of generality we may assume that Σ contains at least one parameter from $\mathbb{N}^+$ and one nominal $\{\ell_i\}$. A set term s is in *negation normal form* (NNF) if $\neg$ occurs in s only in front of set variables and set constants. Each set term can clearly be transformed into an equivalent one in NNF by pushing $\neg$ inwards (e.g., $\neg\exists^{<a}s$ is equivalent to $\forall^{<a}\neg s$). So we may assume all set terms to be in NNF. Distance atoms of $\mathcal{MS}^{\lessgtr}_{\leq}[\mathbb{N}^+]$ can be expressed via nominals and distance quantifiers, e.g.,

$$\begin{array}{lll} \boldsymbol{\delta}(k,\ell) < a & \text{is equivalent to} & k \sqsubseteq \exists^{<a}\{\ell\}, \\ \boldsymbol{\delta}(k,\ell) \geq a & \text{is equivalent to} & k \sqsubseteq \neg\exists^{<a}\{\ell\}. \end{array}$$

So, from now on we will not be considering distance atoms at all. Moreover, without loss of generality we may assume that the language $\mathcal{MS}^{\lessgtr}_{\leq}[\mathbb{N}^+]$ contains no formulas of the form $\ell \sqsubseteq s$. Indeed, every such formula is equivalent to $\{\ell\} \sqcup s \doteq s$. We can also assume that all equalities $s_1 \doteq s_2$ and their negations $s_1 \not\doteq s_2$ are of the form $s \doteq \top$ and $s \not\doteq \bot$, respectively. And by taking for each formula of the form $s \not\doteq \bot$ in Σ a fresh location constant ℓ_s and replacing $s \not\doteq \bot$ with $\ell_s \sqsubseteq s$, we can get rid of all negations of equalities.

To sum up: in what follows (without loss of generality) we confine ourselves to considering those knowledge bases Σ over $\mathbb{N}^+$ that contain only equalities of the form $s \doteq \top$, with s being in NNF.

Let $loc(\Sigma)$ be the set of all location constants occurring in Σ. Define the *closure* $cl(\Sigma)$ of Σ to be the finite set of terms containing $trm(\Sigma)$ and, for all $a \in \mathbb{N}^+$,

$$\{\forall^{<a}s, \forall^{\leq a}s : \exists b \geq a \left(\{\forall^{\leq b}s, \forall^{<b}s\} \cap trm(\Sigma) \neq \emptyset\right)\}.$$

Denote by γ_Σ the maximal parameter which occurs in Σ and set

$$M[\Sigma] = \{1, 2, \ldots, \gamma_\Sigma\}, \quad M[\Sigma]^- = \{a^- : a \in M[\Sigma]\},$$

where the a^- are not in $\mathbb{N}$.

The tableau algorithm we are about to describe operates on *constraint systems* $\mathcal{S} = \langle T, <, L, S\rangle$ for Σ, where $\langle T, <\rangle$ is a forest consisting of the disjoint union of finite intransitive trees $\langle T_\ell, <_\ell\rangle$ with roots $\ell \in loc(\Sigma)$, and L and S are *labelling functions* such that

- with each pair $x < y$, $x, y \in T$, the function L associates a number $L(\{x,y\})$ from $M[\Sigma] \cup M[\Sigma]^-$;
- with each $x \in T$ the function S associates a subset $S(x) \subseteq cl(\Sigma)$.

Roughly, the intuition behind this definition is as follows: T is the set of points of a metric space, L defines distances between some points in this space, and S defines values of set terms from $cl(\Sigma)$.

We start with the *initial constraint system* $\langle T_0, <_0, L_0, S_0\rangle$ for Σ, where $T_0 = loc(\Sigma)$, $<_0 = L_0 = \emptyset$, and $S_0(\ell) = \{\{\ell\}\}$, for every $\ell \in loc(\Sigma)$. Then the initial constraint system is expanded by repeatedly applying (in any order) the *completion rules* in Fig. 1, where blocked nodes are defined as follows.

Denote by $<^+$ the transitive closure of $<$. We say that a node $x \in T$ is *directly blocked by a node* y if $y <^+ x$, $S(x) = S(y)$, but for no distinct $u <^+ x$ and $v <^+ x$ do we have $S(u) = S(v)$. The $<^+$-successors of directly blocked nodes are called *indirectly blocked*. All directly or indirectly blocked nodes comprise the set of *blocked nodes*.

Constraint systems obtained by applying the completion rules to the initial constraint system for Σ will be called *constraint systems for* Σ. Such a constraint system $\mathcal{S}$ is said to be *complete* if one of the following conditions hold:

1. $\mathcal{S}$ contains a node x such that $X_i, \neg X_i \in S(x)$, for some set variable X_i;
2. $\mathcal{S}$ contains a node x, such that $\{\ell_i\}, \neg\{\ell_i\} \in S(x)$ for some location constant $\{\ell_i\}$;
3. none of the rules in Fig. 1 is applicable to $\mathcal{S}$ (in the sense that no application of a rule can change the constraint system).

In cases 1 and 2 we say that the constraint system contains a *clash*.

Theorem 8. *The tableau algorithm, having started from any given constraint system Σ, always terminates* (*i.e., reaches a complete constraint system after finitely many*[2] *steps*)*; it is sound* (*i.e., if there is a complete clash-free constraint system for Σ then Σ is consistent*) *and complete* (*i.e., if Σ is satisfiable, then the tableau algorithm returns a clash-free complete system*)*.*

---

[2] Actually, at most $loc(\Sigma) \cdot |trm(\Sigma)|^{2^{|cl(\Sigma)|}+1}$.

$\mathsf{R}_\sqcap$ If $s_1 \sqcap s_2 \in S(x)$ and x is not indirectly blocked, then set $S(x) := S(x) \cup \{s_1, s_2\}$.

$\mathsf{R}_\sqcup$ If $s_1 \sqcup s_2 \in S(x)$ and x is not indirectly blocked, then set either $S(x) := S(x) \cup \{s_1\}$ or $S(x) := S(x) \cup \{s_2\}$.

R_{nom} If $\{\ell_i\} \in S(x) \cap S(y)$ and neither x nor y is indirectly blocked, then set $S(x) := S(x) \cup S(y)$ and $S(y) := S(x) \cup S(y)$.

$\mathsf{R}_=$ If $s \doteq \top \in \Sigma$ and x is not indirectly blocked, then set $S(x) := S(x) \cup \{s\}$.

$\mathsf{R}_\forall$ If $\forall^{<a} s \in S(x)$ or $\forall^{\leq a} \in S(x)$ and x is not indirectly blocked, then set
$S(x) := S(x) \cup \{s\} \cup \{\forall^{\leq b} s : b < a,\ b \in M[\Sigma]\} \cup \{\forall^{<b} s : b \leq a,\ b \in M[\Sigma]\}$.

$\mathsf{R}_{\forall<}$ If $\forall^{<a} s \in S(x)$, $L(\{y, x\})$ is defined, and x is not indirectly blocked, then:
if $L(\{y, x\}) = a^-$, then set $S(y) := \{s\} \cup S(y)$;
if $L(\{y, x\}) = b < a$, then set $S(y) := \{\forall^{<a-b} s\} \cup S(y)$;
if $L(\{y, x\}) = b^-$ with $b < a$, then set $S(y) := \{\forall^{\leq a-b} s\} \cup S(y)$.

$\mathsf{R}_{\forall\leq}$ If $\forall^{\leq a} s \in S(x)$, $L(\{y, x\}) \in \{b, b^-\}$, and x is not indirectly blocked, then:
if $b = a$, then set $S(y) := \{s\} \cup S(y)$;
if $b < a$, then set $S(y) := \{\forall^{\leq a-b} s\} \cup S(y)$.

$\mathsf{R}_{\exists<}$ If $\exists^{<a} s \in S(x)$, x is not blocked, and $L(\{z, x\}) \notin \{b : b < a\} \cup \{b^- : b \leq a\}$ for any z with $s \in S(z)$, then create a new node $y > x$ and set $L(\{x, y\}) := a^-$ and $S(y) := \{s\}$.

$\mathsf{R}_{\exists\leq}$ If $\exists^{\leq a} s \in S(x)$, x is not blocked, and $L(\{z, x\}) \notin \{b : b \leq a\} \cup \{b^- : b \leq a\}$ for any z with $s \in S(z)$, then create a new node $y > x$ and set $L(\{x, y\}) := a$ and $S(y) := \{s\}$.

Figure 1: Tableau rules.

It is to be noted that the rule $\mathsf{R}_\forall$ is rather inefficient as it stands. Actually, there is no need to add those $\forall^{\leq b} s$ and $\forall^{<b} s$, but then some other rules, as well as the blocking strategy, have to be changed in a subtle way. We show here this 'inefficient variant' to make the tableaux as transparent as possible. The worst case complexity (even of the more efficient variant) of the tableau algorithm does not match the EXPTIME upper bound of Theorem 4, which is actually a feature of all implemented tableau procedures in description logic.

4 Metric logic with numerical variables

Let us consider now the language $\mathcal{MS}^{<}_{\leq}[V]$ which is defined almost precisely as $\mathcal{MS}^{<}_{\leq}[M]$, with the only difference being that instead of parameters (numbers) from M we use *variables* from a countably infinite list $V = \{x_0, x_1 \dots\}$ ranging over $\mathbb{Q}^+$. Given an *assignment* $\mathfrak{a} : V \to \mathbb{Q}^+$ and an $\mathcal{MS}^{<}_{\leq}[V]$-knowledge base Σ, we denote by $\Sigma^{\mathfrak{a}}$ the $\mathcal{MS}^{<}_{\leq}[\mathbb{Q}^+]$-knowledge base which results from Σ by replacing each x_i with $\mathfrak{a}(x_i)$.

Let Γ be a set of *constraints* for variables in V (say, a set of rational linear inequalities over V or polynomial equations). A knowledge base Σ over $\mathcal{MS}^{<}_{\leq}[V]$ is said to *be satisfiable relative to* Γ if there exists an assignment $\mathfrak{a}$ such that

- the $\mathfrak{a}(x_i)$ give a solution to Γ and
- $\Sigma^{\mathfrak{a}}$ is satisfiable.

Here are two simple examples:

Example 9. (1) Γ consists of equalities $x_i = a_i$, where $a_i \in \mathbb{Q}^+$. Then a knowledge base Σ over $\mathcal{MS}^{<}_{\leq}[V]$ is satisfiable relative to Γ iff the knowledge base $\Sigma^{\mathfrak{a}}$ is satisfiable for $\mathfrak{a} : x_i \mapsto a_i$. Thus, in this case $\Sigma^{\mathfrak{a}}$ is a usual $\mathcal{MS}^{<}_{\leq}[\mathbb{Q}^+]$-knowledge base considered in the previous sections.

(2) Γ consists of inequalities $x_i < x_j$, so that no absolute values for distances are fixed, and to satisfy Σ we first have to find a solution (assignment) $\mathfrak{a}$ to Γ and then check whether $\Sigma^{\mathfrak{a}}$ is satisfiable.

To illustrate the expressive power of the new language, we will use the following notation: given set terms s_1 and s_2, denote by $d(s_1, s_2) = x$ the set of two $\mathcal{MS}^{<}_{\leq}[V]$-formulas $s_1 \sqcap \exists^{\leq x} s_2 \not\doteq \bot$ and $s_1 \sqcap \exists^{<x} s_2 \doteq \bot$ (saying that the distance between s_1 and s_2 is x). Then we can state that the distance between Russia and the UK is larger than the distance between Russia and Germany:

$$\Sigma = \{d(\mathsf{Russia}, \mathsf{UK}) = x,\ d(\mathsf{Russia}, \mathsf{Germany}) = y\} \text{ and } \Gamma = \{x > y\}.$$

Incomplete knowledge can also be expressed by using interval constraints $a_1 \leq x \leq a_2$, where x is a variable occurring in the knowledge base.

Of course, as long as algorithmic properties are not relevant, we can consider arbitrarily complex constraints Γ on the variables in V. Our main decidability result, however, holds for the case when Γ consists of a finite set of rational linear inequalities of the form

$$a_1 x_{i_1} + \cdots + a_k x_{i_k} \geq b \quad \text{or} \quad a_1 x_{i_1} + \cdots + a_k x_{i_k} > b.$$

Namely, we have the following:

Theorem 10. *It is decidable whether a given $\mathcal{MS}^{<}_{\leq}[V]$-knowledge base Σ is satisfiable relative to a set of rational linear inequalities.*

The proof of Theorem 10 is based on a (rather straightforward) combination of a linear programming technique with the tableau algorithm presented above. The resulting decision procedure is extremely complex. We conjecture, however, that more sophisticated combinations with lower complexity can be developed. Here is one example:

Theorem 11. *Suppose that Γ consists of the 'pure' strict inequalities*

$$\{0 < x_1, x_1 < x_2, x_2 < x_3, \dots, x_{n-1} < x_n\}$$

and Σ is an $\mathcal{MS}^{\lessgtr}_{\leq}[\{x_1, \dots, x_n\}]$-knowledge base. Then the satisfiability problem for Σ relative to Γ is EXPTIME-complete.

(In fact Σ is satisfiable relative to Γ iff $\Sigma^{\mathfrak{a}}$ is satisfiable, where $\mathfrak{a}(x_1) = 1$ and $\mathfrak{a}(x_i) = 1 + \frac{1}{n+2-i}$, for $1 < i \leq n$.)

5 Conclusion

We have introduced and investigated a 'practical' variant of the distance logics from [Sturm *et al.*, 2000; Kutz *et al.*, 2003]. Our main achievement is a tableau-based decision procedure similar to those implemented in working description logic systems like FACT and RACER, see [Baader *et al.*, 2003]. This similarity gives us some grounds to believe that a system based on this tableau procedure (which is currently being implemented) will be sufficiently fast for 'realistic' applications. As in spatial databases, however, a very important problem is to integrate the distance logic with other KR&R formalisms. Although in the intended applications the extensions of relevant predicates may strongly depend on the distance measure, usually *non-metric* attributes and relations also play a rather significant role (e.g., classifications of proteins, their topological properties). Without an *integration* of formalisms taking care of these different aspects of the problem domain one cannot imagine a system powerful enough for real-world applications. Fortunately, a framework for such an integration is already available. The terms of the language $\mathcal{MS}^{\lessgtr}_{\leq}[\mathbb{Q}^+]$ are constructed from atomic predicates by means of distance and Boolean operators. This makes it possible to treat $\mathcal{MS}^{\lessgtr}_{\leq}[\mathbb{Q}^+]$ as an abstract description system in the sense of [Baader *et al.*, 2002] and combine it in an algorithmically robust way with other effective KR&R formalisms, say, description logics or spatial RCC-8 via E-connections [Kutz *et al.*, 2002] or fusions [Baader *et al.*, 2002].

Acknowledgements

The work on this paper was supported by the DFG-grant Wo 583/3-3 and by U.K. EPSRC grants no. GR/R45369/01 and GR/R42474/01. Thanks are due to Anatoli Degtyarev, Ian Hodkinson, Carsten Lutz, and Oliver Kutz for stimulating discussions and remarks on the preliminary versions of the paper.

References

[Adam and Gangopadhyay, 1997] N. Adam and A. Gangopadhyay. *Database Issues in Geographic Information Systems*. Kluwer, 1997.

[Baader and Sattler, 2000] F. Baader and U. Sattler. Tableau algorithms for description logics. In R. Dyckhoff, editor, *Proceedings of Tableaux 2000*, volume 1847 of *Lecture Notes in Artificial Intelligence*, pages 1–18. Springer, 2000.

[Baader *et al.*, 2002] F. Baader, C. Lutz, H. Sturm, and F. Wolter. Fusions of description logics and abstract description systems. *Journal of Artificial Intelligence Research*, 16:1–58, 2002.

[Baader *et al.*, 2003] F. Baader, D. Calvanese, D. McGuinness, D. Nardi, and P. Patel Schneider, editors. *The Description Logic Handbook: Theory, Implementation, and Applications*. Cambridge University Press, 2003.

[Bozkaya and Özsoyoglu, 1999] T. Bozkaya and Z. Özsoyoglu. Indexing large metric spaces for similarity search queries. *TODS*, 24(3):361–404, 1999.

[Clote and Backofen, 2000] P. Clote and R. Backofen. *Computational Molecular Biology*. John Wiley & Sons, 2000.

[Cohn and Hazarika, 2001] A. Cohn and S. Hazarika. Qualitative spatial representation and reasoning: an overview. *Fundamenta Informaticae*, 43:2–32, 2001.

[Dunham, 2003] M. Dunham. *Data Mining. Introductory and Advanced Topics.* Pearson Education, 2003.

[Hjaltason and Samet, 1999] G. Hjaltason and H. Samet. Distance browsing in spatial databases. *ACM Transactions on Database Systems*, 24(2):265–318, 1999.

[Horrocks, 1998] I. Horrocks. Using an expressive description logic: FACT or fiction? In A. Cohn, L. Schubert, and S. Shapiro, editors, *Proceedings of the 6th International Conference on Principles of Knowledge Representation and Reasoning*, pages 636–647, 1998.

[Kutz *et al.*, 2002] O. Kutz, F. Wolter, and M. Zakharyaschev. Connecting abstract description systems. In D. Fensel, F. Giunchiglia, D. McGuinness, and A. Williams, editors, *Proceedings of the 8th International Conference on Principles of Knowledge Representation and reasoning*, pages 215–226. Morgan Kaufmann, 2002.

[Kutz *et al.*, 2003] O. Kutz, H. Sturm, N.-Y. Suzuki, F. Wolter, and M. Zakharyaschev. Logics of metric spaces. *ACM Transactions on Computational Logic*, 2003. In print.

[Matousek, 2002] J. Matousek. *Lectures on Discrete Geometry*. Springer, 2002.

[Randell *et al.*, 1992] D. Randell, Z. Cui, and A. Cohn. A spatial logic based on regions and connection. In *Proceedings of the 3rd International Conference on Knowledge Representation and Reasoning*, pages 165–176. Morgan Kaufmann, 1992.

[Renz and Nebel, 1999] J. Renz and B. Nebel. On the complexity of qualitative spatial reasoning. *Artificial Intelligence*, 108:69–123, 1999.

[Revesz, 2002] P. Revesz. *Introduction to Constraint Databases*. Springer Verlag, 2002.

[Salton, 1989] G. Salton. *Automated Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*. Addison-Wesley, 1989.

[Setubal and Meidanis, 1997] J. Setubal and J. Meidanis. *Introduction to Computational Molecular Biology*. PWS Publishing Company, 1997.

[Sturm *et al.*, 2000] H. Sturm, N.-Y. Suzuki, F. Wolter, and M. Zakharyaschev. Semi-qualitative reasoning about distances: a preliminary report. In *Logics in Artificial Intelligence. Proceedings of JELIA 2000, Malaga, Spain*, number 1919 in LNAI, pages 37–56. Springer-Verlag, 2000.