Sim-DL: Towards a Semantic Similarity Measurement Theory for the Description Logic $\mathcal{ALCNR}$ in Geographic Information Retrieval

Krzysztof Janowicz

Institute for Geoinformatics
University of Muenster, Germany
janowicz@uni-muenster.de

Abstract. Similarity measurement theories play an increasing role in GIScience and especially in information retrieval and integration. Existing feature and geometric models have proven useful in detecting close but not identical concepts and entities. However, until now none of these theories are able to handle the expressivity of description logics for various reasons and therefore are not applicable to the kind of ontologies usually developed for geographic information systems or the upcoming geospatial semantic web. To close the resulting gap between available similarity theories on the one side and existing ontologies on the other, this paper presents ongoing work to develop a context-aware similarity theory for concepts specified in expressive description logics such as $\mathcal{ALCNR}$.

1 Introduction and Motivation

Within semantic-based geographic information systems and the upcoming geospatial semantic web, ontologies will play a crucial role in semi-automatic information retrieval, integration and concept matching. Two approaches turned out to be useful to support these tasks: subsumption reasoning and similarity measurement.

The idea behind subsumption-based retrieval as described by Lutz & Klien [1] is to rearrange a queried application ontology taking a search concept into account and to return a new taxonomy in which all subconcepts of the injected search phrase satisfy the user's requirements. However, using subsumption reasoning to query knowledge bases forces the user to ensure that the search concept is specified in a way that it is neither too generic and therefore at a top level of the new hierarchy nor too specific to get a sufficient result set. In fact the search concept is a formal description of the minimum characteristics all retrieved concepts need to share. Moreover no measurement structure is provided answering the question *which* of the returned concepts fits best. Yet this is not necessarily a critical point within this approach because all subconcepts at least share the demanded properties. In contrast, similarity computes the degree of overlap between search and compared-to concept and as measurement structure provides a (weak) order. Both characteristics turn out to be useful for information retrieval and matching scenarios: on the one hand the determination of conceptual overlap simplifies

R. Meersman, Z. Tari, P. Herrero et al. (Eds.): OTM Workshops 2006, LNCS 4278, pp. 1681 – 1692, 2006.

phrasing an adequate search concept and on the other hand the results are ordered by their degree of similarity to the *searched* concept. Similarity-based retrieval does not necessarily imply a subsumption relation between search and compared-to concept; in some cases even disjoint concepts may be similar to each other (e.g. Mother, Father). In opposite to subsumption-based retrieval, the search phrase typed into the system is not an artificial construct, but the concept the user is really looking for in the external ontology. The result set describes the measured overlap between compared concept descriptions without presuming that they share a specific property.

In other words the benefits similarity offers during the information retrieval phase, i.e. to deliver a flexible degree of conceptual overlap to a searched concept, stand against shortcomings during the usage of the retrieved information, namely that the results not *necessarily* fit the user's requirements. To make the difference between both approaches more evident (see figure 1), one could imagine a search phrase specified using a shared vocabulary to retrieve all concepts which's instances *overlap* with waterways. In contrast to the subsumption-based approach, similarity measurement will additionally deliver concepts which's instances are located *inside* or *adjacent* to waterways and indicate through a lesser degree of similarity that these concepts are close to, but not identical with the user's intended concept.

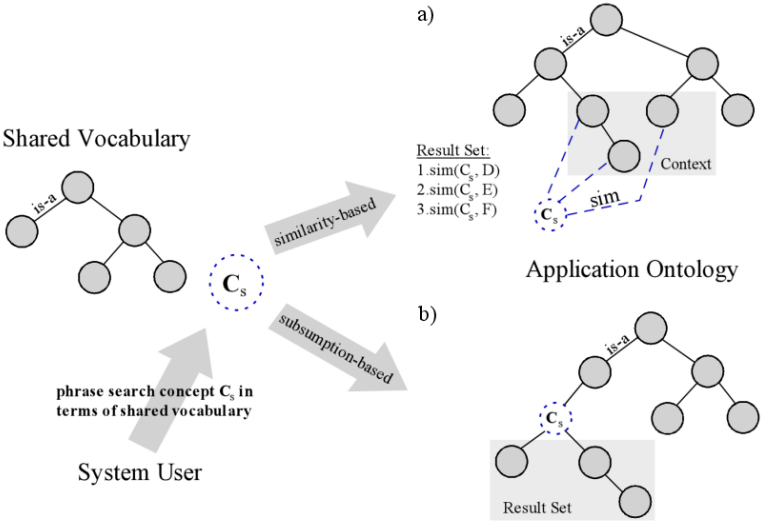

Fig. 1. Similarity (a) and subsumption (b) - based retrieval using a shared vocabulary

Similarity measurement has a long tradition in cognitive science and meanwhile also in computer science and has been applied to various information retrieval scenarios within GIScience for the last years [2-5]. An overview about existing similarity theories, their application areas and characteristics, is out of the scope of this paper and recently presented in [6, 7]. Instead we focus on briefly discussing why yet another similarity theory is necessary. Following the above argumentation, similarity supports users and software agents during information retrieval; however this presumes that the chosen similarity measure does not only satisfy the user's requirements, but also supports the representation language of the inspected ontology. It turns out that, besides the fact that several similarity theories make fundamentally different assumptions about how and *what* is measured (e.g. feature versus geometric model [6]), most of them come with their own proprietary knowledge representation format. In contrast, the majority of ontologies is specified using standardized or commonly agreed logic-based knowledge representation languages and especially various kinds of description logics. Without claiming that logic-based representation is *the* adequate tool for conceptualization and reasoning, we observe a gap between available similarity theories and existing ontologies which opposes a wider application of similarity measures as part of GIS or semantic-enabled web services in general.

In addition, several proprietary knowledge representation formats brought along with existing similarity theories lack of a formal semantic and language constructs proven to be useful for conceptualization, such as disjunction, negation, value and existential restrictions, number restrictions (cardinalities) and roles (binary predicates) in general. This is a crucial point because, at least in computer science, the concepts between which similarity is measured are *representations*[1] of the concepts in our minds. Consequently, the lack of a precise and expressive representation language has impact on the quality of the resulting similarity assessments as discussed in [8] for the lightweight ontology underlying the feature-based MDSM theory [4]. The same arguments hold for geometric approaches to similarity, based on Gärdenfors' [9] idea of conceptual spaces. To integrate relations and hence improve the expressivity of conceptual spaces for similarity measures, Schwering [10] for instance combines the geometric approach with classical network models.

In comparison to independently developed similarity theories for logic-based knowledge representation, such as discussed in [11] for similarity between web services or an approach to measure maximum dissimilarity between concepts represented in $\mathcal{ALC}$ [12], the theory introduced within this paper measures the overall similarity for the high expressive description logic $\mathcal{ALCNR}$. Moreover it supports a (basic) notion of context and conceptual neighborhood models, which are necessary to handle spatial and temporal relations. However, as will be discussed in the future work section, to capture the full extent of geospatial knowledge, even more expressive description logics are necessary [13]. For further work concerning similarity measures between logic-based representations see also [14, 15].

2 Syntax and Semantics of $\mathcal{ALCNR}$

This section gives a brief insight into syntax and semantics of the description logic used as concept representation language within this paper. $\mathcal{ALCNR}$ is an expressive description logic that supports intersection, union, full existential quantification, value restriction, full negation and number restrictions to inductively construct complex concept descriptions out of primitive concepts and roles (binary predicates). In the following sections the letters A and B are used to represent atomic concepts, R and S

[1] This is, at the same time, one of the reasons why we do not claim that the presented similarity theory is necessarily cognitive adequate; however due to lack of space this is not discussed here in detail.

for roles and C and D for complex (composed) concepts, while X and Y denote that a given formula can be applied to all of them. Additional background information about $\mathcal{ALCNR}$ and related description logics is discussed in [16].

Before similarity can be computed, the compared (complex) concepts have to be rephrased to the following $\mathcal{ALCNR}$ disjunctive normal form [17]: A concept description C is in normal form *iff* $C = \top$, $C = \bot$ or $C = C_1 \sqcup \ldots \sqcup C_n$ and each C_i (i= 1,…n) is of the form:

$$C_i := \bigsqcap_{A \in \mathrm{primitive}(C_i)} A \sqcap \bigsqcap_{R \in N_R} \left(\bigsqcap_{C' \in \mathrm{exists}_R(C_i)} (\exists R.C') \sqcap \forall R.\mathrm{forall}_R(C_i) \sqcap \left(\geq \mathrm{min}_R(C_i)R\right) \sqcap \left(\leq \mathrm{max}_R(C_i)R\right) \right)$$

The set *primitive(C)* represents all (negated) primitives (and absurdity) at the top-level of C. N_R is the set of available roles, and $exists_R(C)$ denotes the set of all C' for which there exists $\exists R.C'$ on the top-level of C. $forall_R(C)$ denotes the intersection of concepts ($C_1 \sqcap \ldots \sqcap C_n$) derived by merging all value restriction for the role R ($\forall R.C_i$) on the top level of C. $min_R(C)$ and $max_R(C)$ represent the minimum and maximum cardinalities for the role R on the top-level of C. Complex roles are in conjunctive normal form ($R = R_1 \sqcap \ldots \sqcap R_n$) where each R_i is primitive. Note that the concepts $forall_R(C_i)$ and C' are again in $\mathcal{ALCNR}$ normal form.

Table 1. Syntax and semantics of $\mathcal{ALCNR}$ [16, 17]

Syntax	Semantics	Description
A	$A^{\mathcal{I}} \subseteq \Delta^{\mathcal{I}}$	atomic concept
R	$R^{\mathcal{I}} \subseteq \Delta^{\mathcal{I}} x\, \Delta^{\mathcal{I}}$	atomic role
$\top$	$\Delta^{\mathcal{I}}$	Totality
$\bot$	$\varnothing$	Absurdity
$\neg C$	$\Delta^{\mathcal{I}} \setminus C^{\mathcal{I}}$	full negation
$C \sqcap D$	$C^{\mathcal{I}} \cap D^{\mathcal{I}}$	conjunction (intersection)
$C \sqcup D$	$C^{\mathcal{I}} \cup D^{\mathcal{I}}$	disjunction (union)
$\forall R(C)$	$\{ x \in \Delta^{\mathcal{I}} \mid \forall y. (x, y) \in R^{\mathcal{I}} \rightarrow y \in C^{\mathcal{I}} \}$	value restriction
$\exists R(C)$	$\{ x \in \Delta^{\mathcal{I}} \mid \exists y. (x, y) \in R^{\mathcal{I}} \wedge y \in C^{\mathcal{I}} \}$	existential quantification
$(\leq n\, R)$	$\{ x \in \Delta^{\mathcal{I}} \mid \lvert\{y:(x, y) \in R^{\mathcal{I}} \}\rvert \leq n\}$	number restrictions $(n \in \mathbb{N})$
$(\geq n\, R)$	$\{ x \in \Delta^{\mathcal{I}} \mid \lvert\{y:(x, y) \in R^{\mathcal{I}} \}\rvert \geq n\}$	
$R \sqcap S$	$R^{\mathcal{I}} \cap S^{\mathcal{I}}$	role conjunction

To ensure that the semantic similarity measure is not influenced by syntactic form, rewriting rules as discussed in [17, 18] have to be applied in order to get a canonical representation of the compared concepts. On the one hand these rewriting rules map between equivalent expressions such as $\forall R(\bot)$ and $(\leq 0\ R)$; on the other hand they make sure that only such descriptions are used within concept specifications that (by definition) have impact on the cardinality of the regarded sets.

3 Scenario

This section describes a simplified concept matching scenario to which SIM-DL is applied afterwards. Both the presented scenario and the introduced conceptualizations are intended to briefly demonstrate the abilities *and* shortcomings of the theory instead of trying to develop a meaningful and sound application ontology.

We assume that a European lodging portal on the internet is providing information about accommodations in touristy attractive cities. To avoid maintenance costs, the service provider does not store the information in a local database but dynamically connects to external (geo) web services. However to offer a consistent interface and vocabulary to the portal users, the service provides an own categorization. To do so, the types of accommodations distinguished in the external services have to be mapped to the local terminology. One of the external services, delivering information about accommodations in Amsterdam, provides separate conceptualizations for houseboats and botels[2] while the local knowledge base does not make this distinction.

The task of similarity measurement within this scenario is to propose whether botels should be displayed as hotels or houseboats within the local terminology presented to the system users. The service provider therefore runs a similarity query using the external Botel conceptualization as search phrase (C_s) and Housing as context (C_{lcs}) (see section 4).

Table 2. Conceptualizations for the accommodation service scenario

User defined context (C_{lcs}) and search concept (C_s)
$\mathbf{C_{lcs}} \equiv \text{Housing}$
$\mathbf{C_s} \equiv \text{Boat} \sqcap \text{Hotel (E)} \sqcap \exists\text{inside(Waterway)} \sqcap \forall\text{inside(Waterway)} \sqcap (\leq 1\ \text{inside})$
Concepts/roles defined within the scenario
House $\equiv \text{Building} \sqcap \text{Housing}$
Hotel (E)$\equiv \text{Housing} \sqcap \exists\text{offer(Room)} \sqcap \exists\text{serviceType(Service)}$
Hotel $\equiv \text{House} \sqcap \exists\text{offer(Room)} \sqcap \exists\ \text{serviceType (Service)}$
Youth_Hostel $\equiv \text{Building} \sqcap \text{Housing} \sqcap \exists\text{serviceType (Service} \sqcup \text{SelfService)} \sqcap \exists\text{offer(Room)}$
Botel (E) $\equiv \text{Boat} \sqcap \text{Hotel (E)} \sqcap \exists\text{inside(Waterway)} \sqcap \forall\text{inside(Waterway)} \sqcap (\leq 1\ \text{inside})$
Houseboat $\equiv \text{Boat} \sqcap \text{Housing} \sqcap \exists\text{inside(Waterway)} \sqcap \exists\text{serviceType(SelfService)} \sqcap \forall\text{inside(Waterway)} \sqcap (\leq 1\ \text{inside})$
Cargo_Ship $\equiv \mathit{Boat} \sqcap \mathit{Storage} \sqcap \exists\mathit{inside(Waterway)} \sqcap (\leq 1\ \mathit{inside})$

To avoid debating fundamental difficulties in ontology matching[3], we assume that all services share a common base vocabulary for their primitives (such as described in [1] and depicted in figure 1). Note that the accessory (E) in table 2 denotes concepts from the external service. Moreover for reasons of readability and simplification, the concepts in table 2 are not expanded to their full normal form and those only appearing on the right hand side are assumed to be primitives.

[2] For instance: Hotel Amstel Botel Amsterdam: http://www.amstelbotel.nl/

[3] However we claim that also complex ontology matching tasks benefit from the idea of similarity measurement as demonstrated in [15].

4 SIM-DL

This section stepwise defines a context-aware and directed similarity measure for DL concepts applicable to information retrieval and matching scenarios. As the presented theory combines ideas from feature and network (distance) -based similarity models, conceptual commonalities and differences to existing approaches are pointed out.

In SIM-DL, similarity between concepts in normal form is measured by comparing their $\mathcal{ALCNR}$ descriptions for overlap, where a high level of overlap indicates high similarity and vice versa. As in description logics, (complex) concepts are specified out of primitive concepts and roles using given language constructors (see table 1), similarity is defined as polymorph, binary and real-valued function $\mathbf{X} \times \mathbf{Y} \rightarrow \mathbf{R}[0,1]$ providing implementations for all language constructs offered by the used description logic. The overall similarity between concepts is just the normalized (and weighted) sum of the single similarities calculated for all parts of the concept descriptions. A similarity value of 1 indicates that compared concept descriptions are equal whereas 0 implies total dissimilarity. In the following σ denotes the normalization factor while ω is used to represent weightings. Note however, that for reasons of readability and clarity of the presented equations only the weighting on disjunction level is discussed here in more detail[4]. Additional weightings, responsible for the balance between roles and fillers (range-concepts) or between several kinds of restrictions, are discussed in the further work section.

First of all, to measure similarity it has to be determined which parts of the concept descriptions (specified by the same language constructor) are compared to each other. To do so, the similarity for each element from the Cartesian product X × Y (for a certain constructor on the same level of the normal form) is measured. From the resulting set of tuples, those with the highest similarity value are chosen for further computation; where each X respectively Y is only selected once. In other words, for *each* part of the search concept's description, a counterpart from the compared-to concept's description is chosen in a way that the most similar parts are compared and each expression is only examined once. In the following the set of selected pairs is marked by the letter *S* followed by an abbreviation for the considered constructor.

The presented similarity theory is directed, i.e. asymmetric [4], in a sense that the resulting overall similarity depends on the search direction. Therefore sim(X, Y) is not necessarily equal to sim(Y, X). While each part of the search concept's description is compared to a counterpart from the compared-to concept, some parts of the latter may not be taken into account for comparison. This is always the case if the compared-to concept is specified by more expressions than the search concept. The similarity for these remaining parts is 0 while they do not increase the normalization factor σ. If however the search concept is described by more elements than can be compared, the similarity for these parts is also 0, but σ is increased by 1 for each remaining part. As result the overall similarity is decreased. In other words, if the examined concept in the application ontology is more specific than requested by the user

[4] This weighting is mandatory in a sense that leaving it aside would violate the idea of disjunction; however we do not claim that it is more important for overall similarity than the additional weightings discussed in the further work section.

(via the search concept) this has no impact on the measured overall similarity[5]. On the other side similarity decreases if the user's search concept is more specific than its counterpart in the queried ontology (see also [7]). Defining similarity as ratio between common and available (i.e. shared and distinguishing) parts of considered concept descriptions makes the presented approach comparable to the feature-based MDSM approach [4] and also the Lin's similarity theorem [19]. Note however that in fact SIM-DL compares formal set restrictions, not features (see section 5).

In the previous sections (see also figure 1) context was described as component of similarity-based retrieval. The idea underlying context (first integrated into geospatial similarity measures by Rodríguez & Egenhofer [4]) is on the one hand to determine which parts from the application ontology have to be compared to the search concept and on the other hand to influence the measured similarity making it situation-aware. Within SIM-DL, context is used to *combine* the benefits of subsumption reasoning and similarity-based retrieval (see section 1). Context is defined as a set of concepts from the application ontology that, after reclassification (comparable to the Lutz & Klien approach [1]), are subconcepts of C_{lcs}[6]: (Context = $\{C | C \sqsubseteq C_{lcs}\}$). C_{lcs} itself is specified by the user together with the search concept (C_s) in terms of the shared vocabulary. In other words, context determines the universe of discourse (called application domain in [4]). In the presented accommodation scenario C_{lcs} ensures that all concepts proposed to be similar to Botel at least act as accommodations (subconcepts of Housing). Therefore similarity to cargo ships would not be measured, although they are kinds of boats as well (see table 2).

To compute overall similarity (sim_u) between two concepts C and D in $\mathcal{ALCNR}$ normal form, the similarity between the disjunctions $C_1 \sqcup \ldots \sqcup C_n$ and $D_1 \sqcup \ldots \sqcup D_m$ has to be measured according to equation 1. Simplifying one may argue that this is the maximum similarity occurring during the cross comparison of involved C_i to D_j, which is not the case, because this measure would reflect the maximum possible similarity occurring between certain individuals, but not the overall tendency. Instead, similarity is calculated for each element of SI (the set of tuples (C_i, D_j) chosen for comparison) and weighted (ω) according to their probability. Note that each C_i and D_j is formed by intersection (see $\mathcal{ALCNR}$ normal form) and their similarity is therefore measured by sim_i and described below (see equation 2).

$$sim_u(C, D) = \sum_{(C_i, D_j) \in SI} \omega_{ij} * sim_i(C_i, D_j) \qquad (1)$$

The weighting ω on disjunction level becomes necessary because, in contrast to intersection, each individual that is member of a concept formed by disjunction can be member of all its single concepts or only of some of them. Consequently overall similarity cannot simply be the sum of the similarities between compared C_i and D_j and hence ω acts as adjustable factor for their relative importance. Note that the sum of all ω is always 1. Depending on application area and search strategy, ω can be computed out of the set cardinality (A-Box) of all involved concept on disjunction level, using

[5] Note however, that while directed similarity fits the requirements of information retrieval [7], other tasks may benefit from default similarity [6] which can be achieved by setting the normalization factor (independently of the direction) to the number of selected pairs.

[6] The abbreviation was chosen to refer to the idea of the least common subsumer in DL [16].

probability assumptions (A&T-Box), or from the structure of the examined ontology (T-Box) (see [16] about A-Box and T-Box). The weighted similarities can then be amalgamated the same way as for the intersection constructor.

$$\begin{aligned} &\mathrm{sim}_i(C,D) = \\ &\frac{1}{\sigma}\Big(\sum_{(A,B)\in SP} \mathrm{sim}_p(A,B) + \sum_{(R,S)\in SE} \mathrm{sim}_e(\mathrm{exists}_R(C), \mathrm{exists}_S(D)) + \sum_{(R,S)\in SF} \mathrm{sim}_f(\mathrm{forall}_R(C), \mathrm{forall}_S(D)) \\ &+ \sum_{(R,S)\in SMIN} \mathrm{sim}_m(\mathrm{min}_R(C), \mathrm{min}_S(D_j)) + \sum_{(R,S)\in SMAX} \mathrm{sim}_m(\mathrm{max}_R(C), \mathrm{max}_S(D))\Big) \end{aligned} \tag{2}$$

On the level of intersection, similarity between two (complex) concepts is the sum of similarities derived from mutually comparing their primitive concepts as well as those formed by existential, value and number restrictions/quantification (see equation 2). In addition to the symbols introduced before, the normalization factor σ is defined as the sum of cardinalities derived from the sets of compared tuples (SP, SE, SF, SMIN and SMAX). Consequently the possible results of sim_i range between 0 and 1.

$$\mathrm{sim}_p(A,B) = \frac{|\{C \mid (C \sqsubset A) \sqcap (C \sqsubset B)\}|}{|\{C \mid (C \sqsubset A) \sqcup (C \sqsubset B)\}|} \tag{3}$$

As for primitive concepts[7], similarity cannot be computed as degree of overlap between their descriptions, it has to be determined according to equation 3. SIM-DL considers primitives the more similar, the more common defined concepts both subsume. To be more precise, similarity between primitives is expressed as the ratio between the number of subconcepts of both primitives and the number of subconcepts of one or both of them determined in a given context. However this approach resembles Tversky's ratio model [20] and MDSM [4], it is not asymmetric because this would require a subconcept relationship between A and B or to a common superconcept which is per definition not the case for primitives. Moreover not features in the sense of attributes, functions or parts [4], but subconcepts are compared.

$$\mathrm{sim}_e(\mathrm{exists}_R(C), \mathrm{exists}_S(D)) = \mathrm{sim}_r(R,S) * \sum_{(C'_i, D'_j)\in SE} \mathrm{sim}_u(C'_i, D'_j) \tag{4}$$

$$\mathrm{sim}_f(\mathrm{forall}_R(C), \mathrm{forall}_S(D)) = \mathrm{sim}_r(R,S) * \mathrm{sim}_u(\mathrm{forall}_R(C), \mathrm{forall}_S(D)) \tag{5}$$

$$\mathrm{sim}_m(m_R(C), m_S(D)) = \mathrm{sim}_r(R,S) * \left(1 - \frac{|m_R(C) - m_S(D)|}{m_{RS}(\mathrm{total})}\right) \tag{6}$$

Equation 4, 5 and 6 show how similarity is measured between restrictions and between quantifications. To determine the overlap both parts, the involved roles and the involved fillers (respectively cardinalities) have to be taken into account. Note that *forall*$_R$*(C*$_i$*)* and *C´* are again in normal form (see section 2) while $m_R(C)$ and $m_S(D)$ are numbers restricting the max/min occurrence or the roles R respectively S. In addition to already introduced symbols, sim_r denotes the similarity between roles while *m* acts as abbreviation for *min* respectively *max*, indicating that the same equation is applied

[7] Per definition primitive concepts (also called base symbols) are those which *only* occur on the right hand side of axioms.

for both cases. m_{RS}(total) denotes the highest maximum (respectively minimum) cardinality for the roles R or S in the user defined context. In other words, similarity between number restrictions depends on their relative distance, where m_{RS}(total) reflects the notion of universe in statistics. While the similarity sim_r for primitive roles can be measured following the ideas introduced for primitive concepts (see equation 3), similarity between roles formed by intersection or situated in conceptual neighborhoods is computed according to equation 7 and 8.

$$\mathrm{sim}_{ri}(R,S) = \frac{1}{\sigma} \sum_{(R',S') \in \mathrm{SRI}} \mathrm{sim}_r(R',S') \tag{7}$$

$$\mathrm{nsw}(R,S) = \frac{\mathrm{max\_distance} - \mathrm{edge\_distance}(R,S)}{\mathrm{max\_distance}} \tag{8}$$

$\mathcal{ALCNR}$ supports the composition of roles by intersection, consequently every (complex) role can be expanded to an intersection of primitive roles and hence similarity can be understood as the sum of the similarities for mutually compared (SRI) primitive roles from R and S (sim_{ri}; equation 7). The normalization factor σ becomes necessary to ensure that the derived inter-role similarity ranges between 0 and 1 and can be integrated as part for the similarity measures introduced for restrictions and quantifications.

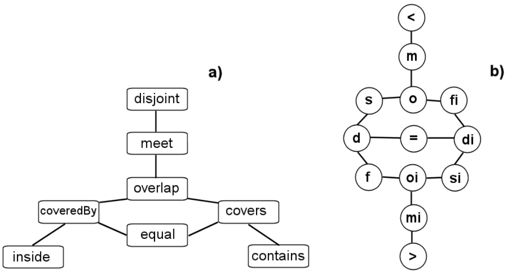

Fig. 2. a) Topological neighborhood [21] b) Temporal-(C) neighborhood [22][8]

Conceptual neighborhoods, such as the topological and temporal depicted in figure 2, are of major importance within GIScience. Per definition one of the benefits of neighborhood models is that they come with an own, straightforward notion of similarity which can be directly integrated as so-called network approach (see also [2, 10]) into SIM-DL. The neighborhood similarity weighting (nsw; equation 8) is defined in terms of edge distance between compared roles (shortest path) and maximum distance within the graph. This approach is comparable to the distance measure by Rada et al. [23]. The edge weightings used for determining distance depend on the neighborhood and even do not need to be symmetric. However they are assumed to be constantly 1 here. Within SIM-DL, nsw can be applied in place of or jointly with sim_{ri} (sim_r). For complex similarity measures between spatial scenes see [21, 24].

[8] [20] introduces three different graphs (A, B and C) for temporal neighborhood, which is not discussed here in detail.

5 Discussion and Further Work

Applying SIM-DL to the accommodation scenario in section 3 yields that Botel is more similar to Houseboat than to Hotel or Youth_Hotel (see table 3) and therefore the service provider can also display botels on the web portal whenever users are searching for houseboats in Amsterdam. Note that in contrast to subsumption-based retrieval the resulting similarities cannot be used to align the concept Botel within the local knowledge base. For instance the definition of Botel is also similar to the local Hotel concept; however a botel is not a building and therefore not a special case (i.e. a subconcept) of a hotel.

Table 3. Measured similarities for the concepts compared in the accommodation scenario

sim(C_s, Hotel)	sim(C_s, Houseboat)	sim(C_s, Youth_Hostel)
0.5	**0.66**	**0.41**

Moreover it has to be emphasized again that similarity in computer science measures overlap between representations. An application ontology about vessels would focus on other aspects then the accommodation ontology and hence the resulting similarities would be different. This is not a shortcoming of the presented theory, but an indicator for the situated nature of conceptualization [25] and hence the importance of context for similarity assessments.

In addition to the results obtained by applying SIM-DL to the accommodation scenario, MDSM[4][9] was also used for comparison. However, due to the different representation languages (and as claimed in section 1) this turns out to be a difficult task. MDSM distinguishes between parts, functions and attributes as features (i.e. characteristics) of the compared conceptualizations. The elements compared by SIM-DL however are formal set restrictions (see semantics of $\mathcal{ALCNR}$; section 2). While primitive concepts can be mapped to features as proposed in [14], the author is very skeptic about applying this method also to (role based) restrictions and quantifications, because features in MSDM are synsets and no notion of fillers or partial matches is defined (see [8]). Specifying the feature *inside* within a concept description in MDSM means that all its instances are inside something, which is not the case for $\forall inside(\top)$ or $\forall inside(\text{Waterway})$. Moreover basing on Tversky's ratio model, MDSM regards concepts as bags of features and therefore it is no clear how to integrate disjunction ($\sqcup$) into this approach.

Nevertheless the comparison between SIM-DL and MDSM points out an interesting aspect of the presented approach: If several constructors (for the same role) are necessary to restrict an intended set (such as for *inside* in table 2) this has multiple impact on the measured similarity[10]. This is not problematic from a set theoretic point of view, but not the way humans think about similarity (see also remark in section 1).

Although by integrating measures for role-based constructors, role intersection and neighborhoods, SIM-DL meets the demands claimed for modern inter-concept

[9] The same remarks also count for Tversky's ratio model.

[10] Note however that the redundant definitions for *inside* in table 2 are captured by the rewriting rules for the canonical normal form.

similarity theories [4, 7, 8], the presented approach is still in progress and a lot of work remains to be done: In addition to the weighting introduced on disjunction level, further weightings should be integrated into SIM-DL to balance the importance between roles and fillers for existential quantification (or value restrictions). However, defining sim_e as weighted sum of role and filler similarity raises the question how such weightings should be derived and whether the role or filler part of an existential quantification is more or less important for sim_e (and therefore overall similarity). Additional weightings should also determine the relative importance of language constructs. In terms of the presented scenario, the level of information about botels provided by ∃inside(Waterway) is higher than (≤ 1 inside), because the last mentioned expression only stats that a botel is at most inside (2D) one thing. Further work has to examine whether these weightings can be (semi)-automatically derived from the context, the kind of chosen description logic and canonical form. The integration of *Inference based Information Value* [11] and other information-theoretic approaches [19] into SIM-DL seems to be a promising approach.

Moreover, until now SIM-DL does not support cyclic concept definitions. To overcome this shortage, techniques such as fixpoint semantics have to be integrated into the theory. Additional work is necessary to develop similarity theories for even more expressive description logics (such as $\mathcal{ALCRP(D)}$ [13]), especially focusing on full qualifying number restrictions and concrete domains. In terms of the presented scenario this would allow to express *how* near something is to a waterway instead of merely distinguishing between inside, meet and overlap. Finally computation time is a critical aspect (especially for the Cartesian products) to be examined in more detail.

References

1. Lutz, M. and E. Klien, *Ontology-Based Retrieval of Geographic Information.* International Journal of Geographical Information Science, 2006. **20**(3): p. 233-260.
2. Janowicz, K., *Towards a Similarity-Based Identity Assumption Service for Historical Places*, in *Geographic Information Science - Fourth International Conference, GIScience 2006.Lecture Notes in Computer Science 4197*, M. Raubal, et al., Editors. forthcoming 2006, Springer: Berlin, Germany.
3. Raubal, M., *Formalizing Conceptual Spaces*, in *Formal Ontology in Information Systems, Proceedings of the Third International Conference (FOIS 2004)*, A. Varzi and L. Vieu, Editors. 2004, IOS Press: Amsterdam, NL. p. 153-164.
4. Rodríguez, A.M. and M.J. Egenhofer, *Comparing Geospatial Entity Classes: An Asymmetric and Context-Dependent Similarity Measure.* International Journal of Geographical Information Science, 2004. **18**(3): p. 229-256.
5. Schwering, A. and M. Raubal, *Measuring Semantic Similarity between Geospatial Conceptual Regions*, in *First International Conference on GeoSpatial Semantics, GeoS 2005, Mexico City, Mexico.* 2005, Springer-Verlag: Berlin. p. 90-106.
6. Goldstone, R. and J. Son, *Similarity*, in *Cambridge Handbook of Thinking and Reasoning*, K. Holyoak and R. Morrison, Editors. 2004, Cambridge University Press: Cambridge.
7. Schwering, A., *Semantic Similarity Measurement including Spatial Relations for Semantic Information Retrieval of Geo-Spatial Data.* (submitted 2006): Institute for Geoinformatics, University of Münster, Germany, PhD Thesis.

8. Janowicz, K., *Extending Semantic Similarity Measurement by Thematic Roles*, in *First International Conference on GeoSpatial Semantics, GeoS 2005, Mexico City, Mexico.*2005, Springer Verlag: Berlin. p. 137-152.
9. Gärdenfors, P., *Conceptual Spaces - The Geometry of Thought*. 2000, Cambridge, MA: Bradford Books, MIT Press. 307.
10. Schwering, A. *Hybrid Model for Semantic Similarity Measurement*. in *4th International Conference on Ontologies, DataBases, and Applications of Semantics (ODBASE05)*. 2005. Agia Napa, Cyprus: Springer.
11. Hau, J., W. Lee, and J. Darlington. *A Semantic Similarity Measure for Semantic Web Services*. in *Web Service Semantics Workshop 2005 at WWW2005*. 2005. Chiba, Japan.
12. d'Amato, C., N. Fanizzi, and F. Esposito. *A Semantic Dissimilarity Measure for Concept Descriptions in Ontological Knowledge Bases* in *The Second International Workshop on Knowledge Discovery and Ontologies*. 2005. Porto, Portugal.
13. Möller, R., *Expressive Description Logics: Foundations for Practical Applications*. Habilitation Thesis. 2001, University of Hamburg, Computer Science Department, Germany.
14. Borgida, A., T.J. Walsh, and H. Hirsh. *Towards Measuring Similarity in Description Logics*. in *International Workshop on Description Logics (DL2005)*. 2005. Edinburgh, Scotland.
15. Ehrig, M., et al. *Similarity for Ontologies - A Comprehensive Framework*. in *13th European Conference on Information Systems*. 2005. Regensburg, Germany.
16. Baader, F. and W. Nutt, *Basic Description Logics*, in *The Description Logic Handbook*, D.C. F. Baader, D.L. McGuinness, D. Nardi, P.F. Patel-Schneider, Editor. 2002, Cambridge University Press: Cambridge p. 47-100.
17. Brandt, S., R. Küsters, and A.Y. Turhan, *Approximating ALCN-Concept Descriptions*, in *Proceedings of the 2002 International Workshop on Description Logics*. 2002.
18. Molitor, R., *Structural Subsumption for ALN. LTCS-Report 98-03, LuFG Theoretical Computer Science, RWTH Aachen, Germany.* 1998.
19. Lin, D., *An information-theoretic definition of similarity*, in *Proceedings of the Fifteenth International Conference on Machine Learning*. 1998, Morgan Kaufmann, San Francisco, CA. p. 296-304.
20. Tversky, A., *Features of Similarity.* Psychological Review, 1977. **84**(4): p. 327-352.
21. Bruns, T.H. and M.J. Egenhofer, *Similarity of Spatial Scenes*, in *Seventh International Symposium on Spatial Data Handling (SDH '96)*, M.-J. Kraak and M. Molenaar, Editors. 1996: Delft, Netherlands. p. 31-42.
22. Freksa, C., *Temporal Reasoning Based on Semi-Intervals.* Artificial Intelligence, 1992. **54**(1): p. 199-227.
23. Rada, R., et al., *Development and Application of a Metric on Semantic Nets.* IEEE Transaction on Systems, Man, and Cybernetics, 1989. **19**(1): p. 17-30.
24. Li, B. and F.T. Fonseca, *TDD - A Comprehensive Model for Qualitative Spatial Similarity Assessment.* Spatial Cognition and Computation, 2006. **6**(1): p. 31-62.
25. Barsalou, L., *Situated simulation in the human conceptual system.* Language and Cognitive Processes, 2003. **5**(6): p. 513-562.