Similarity-based Information Retrieval and its Role within Spatial Data Infrastructures

Krzysztof Janowicz[1], Marc Wilkes[1], Michael Lutz[2]

[1] Institute for Geoinformatics, University of Muenster, Germany
janowicz|marc.wilkes@uni-muenster.de
[2] European Commission – Joint Research Centre
Institute for Environment and Sustainability, Ispra, Italy
michael.lutz@jrc.it

Abstract. While similarity has gained in importance in research about information retrieval on the (geospatial) semantic Web, information retrieval paradigms and their integration into existing spatial data infrastructures have not been examined in detail so far. In this paper, intensional and extensional paradigms for similarity-based information retrieval are introduced. The differences between these paradigms with respect to the query and results are pointed out. Web user interfaces implementing two of these paradigms are presented, and steps towards the integration of the SIM-DL similarity theory into a spatial data infrastructure are discussed. Remaining difficulties are highlighted and directions of further work are given.

1 Introduction and Motivation

Semantics-based information retrieval [1, 2] plays an increasing role in GIScience and research on the geospatial semantic Web. In general, two approaches can be distinguished, those based on classical subsumption reasoning and those based on so-called non-standard inference techniques [3, 4] – similarity being one of them. While the number of similarity theories for information retrieval is increasing, the number of real geo-application is still low. So far, the most prominent reason was that existing theories were not able to handle the expressivity of description logics used by ontologies on the (geospatial) semantic Web. Recently developed theories [5–8] bear the potential to close this gap, which moves the focus towards new challenges. This paper addresses one of these challenges: How can similarity be integrated within existing spatial data infrastructures (SDIs) to support users during information retrieval?

The paper is based on the SIM-DL similarity theory [7], which is introduced in section 2. This section also describes relevant background in the areas of similarity measurement, description logics, and SDIs. Sections 3 and 4 present the two main contributions of the paper. In section 3, we present a classification of semantics-based information retrieval paradigms. We investigate the differences between subsumption and similarity-based approaches in terms of how a query is phrased and which results can be expected. The examined paradigms are grouped into intensional and extensional approaches (and combinations of both). In section 4, we discuss how the similarity-based paradigms can be implemented using a SDI and what difficulties have to be overcome. Section 5 concludes the paper and points out directions for future research.

In order to illustrate our work, the following scenario is used: A tourist, Nicole, is planing a trip to Utah, US. As a passionate canoeist she would like to spend some time canoeing. Since she has never been to Utah before, she does not know any waterbodies that might please her. However, she does remember some rivers and lakes that she canoed a few years ago while visiting friends in Canada. Therefore, she decides to search for this kind of features browsing services on the Web. We assume that these services are part of a SDI that also includes a catalogue service providing information about available geographic feature types and a web similarity service (WSS) based on the SIM-DL server [7].

2 Related Work

This section introduces related work. The objectives and basic components of SDIs are presented in section 2.1. Sections 2.2 and 2.3 describe similarity measurement and description logics, respectively. These provide the basis for the SIM-DL similarity theory introduced in section 2.4.

2.1 Spatial Data Infrastructures

The main goal of SDIs is to offer access to distributed data sources. The development of interoperability specifications – and here most prominently the work within the Open Geospatial Consortium[3] (OGC) – has created a technology evolution that moves from standalone GIS applications towards a more loosely coupled and distributed model based on self-contained, specialized, and interoperable (Web) services [9]. In such an infrastructure, where resources are distributed and controlled by different organizations, catalogue services provide a means for describing the services' locations and capabilities. They store metadata and support users in discovering and using these resources. The OGC has specified a catalogue service for the Web (CS-W) and related metadata profiles.

While thus the SDI concept promises an efficient sharing and reuse of geographic data among heterogeneous user groups [9, 10], most existing SDIs are still at an early stage in their development. Many just offer geoportals that integrate on-line map viewers and catalogue services for their data holdings [11, 12].

In this work, we focus on geographic data provided through the Web feature service interface (WFS) and organized into geographic feature types. In SDIs, metadata about feature types can be stored in so-called *feature catalogues* [13]. Feature catalogues define the types of features, their operations, attributes, and

[3] The OpenGIS standards and specifications are available from http://www.opengeospatial.org/standards/.

associations. Their goal is to provide a better understanding of geographic data in order to enable users to judge whether the data fits their purpose. An implementation of a feature type catalogue as an extension package for the ebRIM Profile of the CS-W has been proposed [14].

2.2 Similarity Measurement

Similarity has its origin in cognitive science and was established as a theory to investigate how entities are grouped into categories, and why some categories (and their members) are comparable while others are not [15, 16]. *Semantic* similarity measures the proximity of meanings as opposed to purely structural comparison. While entities can be expressed in terms of attributes, the representation of concepts[4] is more complex. In dependence of the (computational) characteristics of the representation language, concepts are specified as unstructured bags of features[5], regions in a multidimensional space, or set-restrictions specified using various kinds of description logics. As the computational concepts are models of concepts in human minds, similarity depends on what is said (in terms of representation) about these concepts. Context is the next big challenge for similarity research. In most cases, meaningful notions of similarity cannot be established without defining in respect to what similarity is measured [16, 18–20].

Similarity-based information retrieval plays an increasing role in GIScience. Based on Tversky's feature model [17], Rodríguez and Egenhofer [21] developed the Matching Distance Similarity Measure which offers a basic context theory, feature weights, and asymmetry, while Raubal [22] proposed geometric similarity measures based on conceptual spaces. Several measures [5–8] were developed to close the gap between ontologies specified in description logics and similarity theories which had not been able to cope with the expressivity of these languages. Other similarity theories [23, 24] have been established to determine the similarity between spatial scenes and also investigate how to use similarity within spatial queries [24]. The ConceptVISTA [25] ontology management and visualization toolkit uses similarity for knowledge retrieval and organization.

2.3 Description Logics and Inference

Description logics (DL) are a family of knowledge representation languages used to model concepts and individuals within a knowledge base. A knowledge base consists of a TBox which contains the terminology, i.e., the concepts within a given domain, and an ABox which stores assertions (about named individuals). A DL system offers services to reason about the content of a knowledge base. Standard reasoning services include satisfiability, subsumption, and instance checking. The computation of the most specific concept (MSC) for an

[4] The term *concept* is used in this paper for the ontological representation of geographic feature types and should not be confused with the concepts in human minds.

[5] In the sense of concept characteristics, see [17].

individual and the least common subsumer (LCS) of several concepts are so-called non-standard reasoning services [4]. Computing the MSC of an individual yields the least concept that the individual is an instance of. Accordingly, the LCS of some concepts is the least concept that subsumes all of them, i.e., there is no subconcept of the LCS which is a superconcept of all these concepts.

Computing the MSC and LCS can serve a variety of purposes. A top-down approach for constructing knowledge bases is not always feasible, since not all relevant concepts might be known beforehand. Instead, one could only specify the building blocks in the TBox and then introduce typical examples as individuals in the ABox. By computing the MSC (and LSC) for these individuals, more complex concepts can be added to the TBox. A semantics-based retrieval approach based on computing the LCS has been proposed by Möller et al [3].

2.4 SIM-DL Similarity Theory and Implementation

While the previous sections focus on similarity in general, this section gives an insight into a similarity theory (and its implementation) which measures similarity between concepts specified in expressive description logics.

SIM-DL [6, 7] is an asymmetric and context-aware similarity measurement theory used for information retrieval. It compares a search concept C_s with a set of target concepts $\{C_{t_1}, ..., C_{t_m}\}$ from an ontology (or several ontologies using a shared top-level ontology). The concepts themselves can be specified using various kinds of expressive description logics. The compared-to target concepts can be either selected by hand, or derived from the so-called context of discourse $\mathcal{C}_d$ [6, 19, 18], i.e., a subset of the ontology, also referred to as the domain of application [21]. It is defined as the set of concepts which are subsumed by the context concept C_c ($\mathcal{C}_d = \{C_t | C_t \sqsubseteq C_c\}$). Hence, each (named) concept $C \in \mathcal{C}_d$ is a target concept for which the similarity $sim(C_s, C_t)$ is computed. Besides cutting out the set of compared concepts, $\mathcal{C}_d$ also influences the resulting similarities (see [6, 7] for details). With respect to the canoeing scenario this means that the similarity between the concepts $River$ and $Canal$ also depends on whether C_c is set to $Watercourse$ or the more general $Waterbody$ (see figure 1). Up to now, the user has to specify the context concept manually. Information retrieval paradigms overcoming this restriction are discussed in section 3. SIM-DL offers an extended context model, but we focus on $\mathcal{C}_d$ here (see [18]).

SIM-DL compares two DL concepts in canonical form [6, 26] by measuring the degree of overlap between their definitions. A high level of overlap indicates a high similarity and vice versa. DL concepts are specified by applying language constructors, such as intersection or existential quantification, to primitive concepts and roles – hence forming complex concepts. Consequently, similarity is defined as a binary and real-valued function $C_s \times C_t \rightarrow$ R[0,1] providing implementations for all language constructs offered by the used description logics. Finally, the overall similarity between concepts is the normalized (and weighted) sum of the single similarities calculated for all parts of the concept definitions. A similarity value of 1 indicates that the compared concepts cannot be differentiated, whereas 0 indicates that they are not similar at all. SIM-DL is an asymmetric measure, i.e., the similarity $sim(C_s, C_t)$ is not necessarily equal to $sim(C_t, C_s)$. Therefore, the comparison of two concepts does not only depend on their descriptors, but also on the direction in which both are compared. In case of concepts composed by disjunction, SIM-DL distinguishes between two similarity modes, the maximum similarity and the average similarity. In the first case, similarity depends on the most similar concept that is part of the disjunction. In the second case, similarity is defined as the average of all involved concepts. While this distinction is a consequence of the used representation language and the definition of similarity functions in SIM-DL [6, 7], we will discuss its importance for information retrieval in section 3.

A single similarity value computed between two concepts hides most of the important information. It does not answer the question whether there are more or less similar target concepts in the examined ontology. It is not sufficient to know that possible similarity values range from 0 to 1 as long as their distribution is unclear [7, 18]. Besides these interpretation problems, isolated comparison puts too much stress on the concrete similarity value. It is hard to argue that and why the result is (cognitively) plausible without other reference values [18]. Consequently, SIM-DL focuses on similarity rankings. The result of a similarity query is an ordered list with descending similarity values. SIM-DL, supports various result representations including font-size scaling or categorization.

The SIM-DL theory is implemented as semantic similarity service (called WSS here). The current (beta) release[6] 2.2 supports subsumption reasoning and similarity measurement up to $\mathcal{ALCHQ}$, as well as MCS and LCS computation (up to $\mathcal{ALE}$). More details on the SIM-DL implementation and a similarity plug-in to the Protégé ontology editor are given by Janowicz et al. [7]. The extensions to the description logics communication interface DIG, necessary to integrate the WSS within the Semantic Web are discussed by Wilkes and Janowicz [27].

3 Intensional and Extensional Retrieval

In this section, we will introduce paradigms for similarity-based information retrieval and group them into intensional and extensional approaches (and combinations of both). We will motivate the need for similarity by contrasting it to approaches purely based on subsumption reasoning. To illustrate the differences, we define two concepts, the *intended concept* C_i, which represents exactly the information the user is looking for, and the *search concept* C_s, which is the concept actually used in the search. We make this distinction, because we assume that in most cases the intended concept is not part of the queried ontology and has to be approximated by the search concept. The result of a query, i.e., the concepts or individuals returned to the user, is the better the more accurate C_s approximates C_i. In the ideal case, C_s would be equal to C_i. If ${C_s}^{\mathcal{I}} \subset {C_i}^{\mathcal{I}}$, all instances of C_s fulfill the user's requirements. In contrast, if ${C_s}^{\mathcal{I}} \supset {C_i}^{\mathcal{I}}$, some instances of C_s might not fulfill the user's requirements. Consequently, a user

[6] The release can be downloaded at http://sim-dl.sourceforge.net/.

interface should support users in defining appropriate search concepts. We further introduce the notions of target concepts C_t and instances I_t, which will be used in our descriptions of similarity-based search paradigms.

In the following, we assume that the information the user is searching for is represented as individuals and concepts within an ontology, and that she uses these artifacts as the basis for her search. How this representation can be mapped to features and feature types provided in a SDI is presented in section 4.

To illustrate the different approaches, we will use an excerpt of a hydrology ontology (figure 1). For readability, the figure shows a simplified representation of the ontology only including its is-a relations; more expressive DL statements (including concept defintions based on roles such as $hasFlowVelocity$ or $hasOwnership$) are not depicted. Note that such expressive statements can also be used in the proposed paradigms and SIM-DL [7].

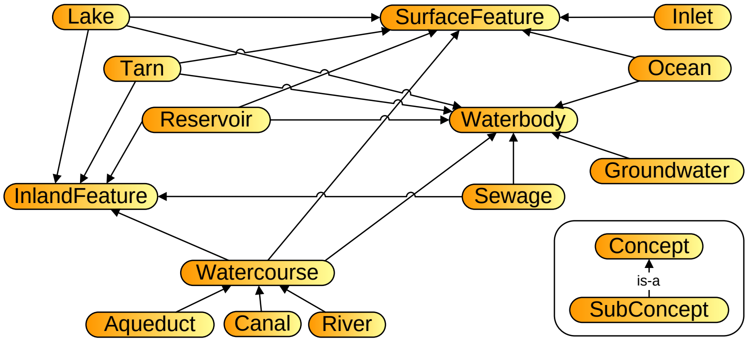


Fig. 1. Fragment of an hydrology ontology.

Referring to the canoeing scenario, Nicole is searching for the ad-hoc category [28] of *canoe-able waterbodies* (C_i). As a corresponding concept is not available in the ontology, she needs to define a query using existing concepts. We assume a graded structure [28, 29] for such ad-hoc categories, i.e., there are more typical and less typical canoe-able waterbodies (on concept and instance level).

3.1 Subsumption-based Retrieval

In terms of subsumption-based retrieval and as depicted in figure 2, the following (intensional) approaches for deriving a query, i.e., the search concept, can be distinguished (see also [1]):

a) Search concept is a subconcept of C_i. The user can specify a particular search concept (e.g., *River*) that is known to be a subconcept of C_i. Consequently, while all individuals of C_s also satisfy the requirements for C_i, many appropriate individuals will not be captured as ${C_s}^{\mathcal{I}} \subset {C_i}^{\mathcal{I}}$.

- **b) Search concept is a superconcept of C_i.** The user can specify a particular search concept (e.g., $Waterbody$) that is known to be a superconcept of C_i. Consequently, while all individuals of C_i will be returned, many inappropriate individuals (e.g., instances of $Tarn$, $Ocean$, $Sewage$, or $Groundwater$) will also be captured as ${C_s}^{\mathcal{I}} \supset {C_i}^{\mathcal{I}}$.
- **c) Search concept is defined by conjunction.** The user can try to build a more appropriate search concept using a conjunction of existing named concepts (e.g., $SurfaceFeature \sqcap InlandFeature \sqcap Waterbody$). First, this would require a complex user interface. Second, without a detailed knowledge of the examined ontology, the relation between C_i and C_s remains unclear to the user. Additionally, there is a high likelihood to get an empty result set (because of the conjunction constructor). Thus, such an approach is more suitable for ontology engineers than for end-users.
- **d) Search concept is defined by disjunction.** The user can select several concepts known to be subconcepts of C_i, hence forming C_s as disjunction of those concepts (e.g., $Canal \sqcup River \sqcup Lake \sqcup Reservoir \sqcup Inlet$). While this would require a complex user interface and is time consuming, it would result (if the concepts are carefully selected) in a good approximation of C_i.

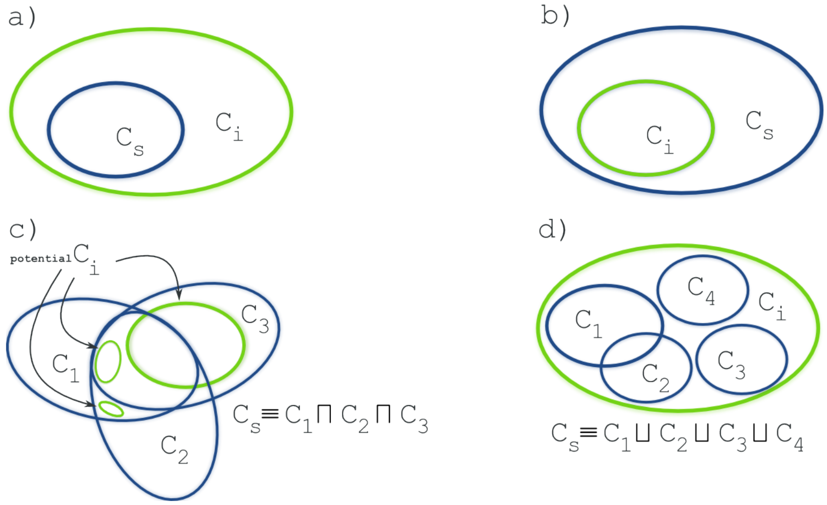


Fig. 2. Subsumption-based information retrieval paradigms.

3.2 Similarity-based Retrieval

In terms of similarity-based retrieval and as depicted in figure 3, the following approaches for deriving a query can be distinguished.

Intensional Paradigm The intensional information retrieval paradigm exclusively relies on concept descriptions to reason about similarity, i.e., no individuals are taken into account. Consequently, a concept ranking is returned to the user.

e) Prototypical search concept. The user can specify a prototypical search concept C_s (e.g., *River*) such as described in approach *a* and define a context concept C_c (e.g., $Waterbody$) in addition, which is known to be a superconcept of C_i (as in *b*)[7]. All subconcepts of C_c, called target concepts C_t here, are compared for similarity to C_s; see section 2.4. As we assume a graded structure, a decreasing similarity to the search concept is interpreted as less intended concept. Such a ranking can already be delivered back to the user [18, 30]. While this approach is comparable to a combination of *a* and *b*, the ranking is a major advantage and supports the user in generating queries such as in case *d*, without requiring a detailed insight into the underlying conceptualizations. While the selection of a prototypical concept is less difficult, finding an appropriate context concept manually is more difficult. As each concept returned within the ranking is a subconcept of C_c, it follows that if C_c does not capture the minimum characteristics of C_i , inappropriate concepts may be returned (however, they would have a low position in the ranking due to their low similarity to the search concept). Subconcepts of C_c whose similarity is 0 (or below a pre-defined threshold) are *not* part of the ranking. Each concept C_t from the ranking satisfies the more probable the user's requirements, the higher its similarity is to C_s.
Additionally, one could automatically generate a new search concept C_{sn} as a disjunction of the returned similar concepts (above a certain threshold) which would be compareable to approach *d* without that the user needs to find and select those concepts by hand. Note that one cannot think of the instances as members of a fuzzy set with the similarity of their concepts (to C_i) as the degree of membership. In SIM-DL (and most related measures), inter-concept similarity cannot be directly mapped to inter-instance similarity (i.e., a similarity $sim(River, Canal)$ of 0.76 does not imply that the similarity between all rivers to all canals is 0.76).

Extensional Paradigm The extensional information retrieval paradigm is a *query-by-example*, and hence relies exclusively on individuals to reason about similarity. A concept (the LCS) computed from the set of examples, called reference individuals here, is used to pre-select the compared-to target individuals. Consequently, the user's query is answered by returning ranked individuals.

f) Reference individuals (for individual similarity). The user can specify a set of reference individuals $\{I_{r_1}, ..., I_{r_n}\}$ (e.g., particular rivers and lakes). The retrieval of target individuals can then be subdivided into three steps. First, the most specific concept MSC_{r_i} for each reference individual

[7] Approach *e* can also be modified to take concepts formed by disjunction or conjunction (such as in *c* and *d*, repsectively) into account.

is determined. The second step is the computation of the least common subsumer for $\{MSC_{r_1}, ..., MSC_{r_n}\}$. The LCS comprises those characteristics that are common to all MSCs, and is therefore used as the context concept C_c[8]. Finally, retrieving the instances of C_c yields the target individuals $\{I_{t_1}, ..., I_{t_m}\}$ (particular rivers, lakes, reservoirs, canals, etc.). Since all target individuals instantiate C_c, they share the same characteristics common to all reference individuals. Now, similarity is used to account for those characteristics that differ among the reference individuals. For example, a characteristic that is common to $\{I_{r_1}, ..., I_{r_{n-1}}\}$, but does not apply to I_{r_n}, is not captured by C_c, and is therefore no requirement for a target individual. Nevertheless, a target individual that shares that characteristic with $n-1$ reference individuals might be more relevant than target individuals lacking that property. The overall relevance of a target individual can be determined by comparing it to each of the reference individuals and combining (e.g., averaging) the resulting similarities. The result returned to the user is a ranking of target individuals illustrating their similarity to the reference individuals.

Recapitulating, the reference individuals are an extensional way to approximate C_i, and at the same time their common characteristics are used as context concept. The set of target individuals might capture inappropriate individuals as $\{I_{r_1}, ..., I_{r_n}\} \subset C_i^{\mathcal{I}} \subseteq \{I_t | I_t \in C_c \ and \ I_t \notin \{I_{r_1}, ..., I_{r_n}\}\}$ holds. Due to the similarity ranking of target individuals, this drawback is compensated. The higher a target individual is ranked the more likely it is within $C_i^{\mathcal{I}}$. In contrast to paradigm *e*, one could also think of the returned ranking as a fuzzy set by replacing the crisp membership (or instance-of) relation with its counterpart from fuzzy sets theory[9]. Then, the degree of membership of a certain (target) individual is given by its similarity value.

Combinations of the Intensional and Extensional Paradigm A combination of the intensional and extensional paradigms reduces the difficulties in selecting appropriate search and context concepts by allowing for the selection of reference individuals (however, both paradigms return (similar) concepts).

g) Prototypical search concept *and* reference individuals. The user can specify a search concept (e.g., *River*) and a set of reference individuals (e.g., waterbodies Nicole had canoed before). This is a combination of the paradigms *e* and *f*, where the search concept is used for comparison and the least common subsumer of the reference individuals is used as context concept to define the context of discourse. The result is a concept ranking

[8] Consider the following example: $MSC_{r_1} \equiv River \sqcap \exists hasFlowVelocity.Velocity \sqcap \exists hasOwnership.Public$ and $MSC_{r_2} \equiv Lake \sqcap \exists hasOwnership.Public$ results in the LCS $C_c \equiv SurfaceFeature \sqcap Waterbody \sqcap InlandFeature \sqcap \exists hasOwnership.Public$. The restriction *hasOwnership.Public* is common to both MSCs, and although *Lake* and *River* do not match, their common superconcepts $SurfaceFeature$, $Waterbody$, and $InlandFeature$ are considered by the LCS.

[9] See Cross and Sudkamp [31] for an overview on fuzzy sets and similarity.

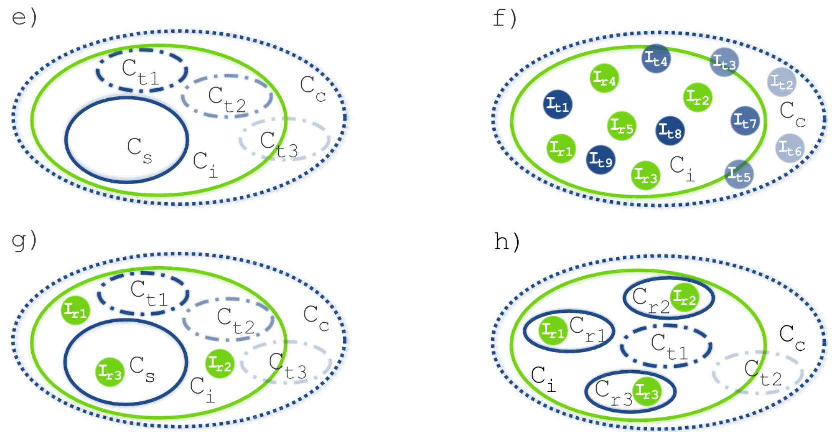


Fig. 3. Similarity-based information retrieval paradigms.

such as in *e*. The advantage is that the user does not need to specify the context concept manually. However, the distinction between search concept and reference individuals may be difficult to explain to the user.

h) Reference individuals (for concept similarity). The user can specify a set of reference individuals. Their LCS is computed and acts as context concept. All resulting target concepts ($C_{t_j} \sqsubseteq C_c$) are compared to the concepts (C_{r_i}) the reference individuals are instances of. For example, if one reference instance is a *Lake*, one a *River* and a third a *Canal*, the LCS would be $SurfaceFeature \sqcap Waterbody \sqcap InlandFeature$ and the target concepts all concepts from our ontology fragment except *Ocean*, *Inlet*, *Groundwater* and *Sewage*. Each target concept is compared to each C_r (*Lake*, *Canal*, and *River* in our example); note that each C_r is a subconcept of C_c, and therefore a target concept itself. This raises the question of how the resulting ranking should be generated. The similarity modes introduced for SIM-DL allow for two different solutions. Out of the user's reference individuals a search concept can be defined as disjunction of the respective reference concepts. Next, either the average or maximum similarity mode can be used to compute the similarity $sim(C_s, C_{t_j})$. In the first case, each C_t is compared to all C_r and the average of its similarity values is used for the ranking. In the second case, the ranking depends on the highest similarity value to one of the C_r. With respect to our example and the average mode, *River*, *Canal*, and *IrrigationCanal* would occupy a higher position in the ranking, followed by *Lake* (see figure 5). This is due to the two watercourses selected as reference individuals.

As in *g*, this approach is also a combination of *e* and *f*, but does not require the manual definition of the search concept. The user only needs to specify reference individuals while C_s and C_c are computed automatically.

3.3 Summary

In contrast to the subsumption-based approaches, similarity supports the user in phrasing queries and delivers a ranking to help the user in judging how well returned individuals or concepts fit her requirements. In the cases *a* and *d*, it is guaranteed that all returned concepts are subconcepts of the intended concepts, i.e., fulfill the user's requirements – this is not the case for similarity-based retrieval in general. To overcome this difficulty the context concept is introduced to capture the minimal characteristics and only its subconcepts are compared for similarity. The approaches *e*–*h* offer different solutions on how to reduce the effort of phrasing the search and the context concept, and automate these steps. The question which of the proposed paradigms fits best depends on the application area. In general, the focus of similarity is more to facilitate the navigation and browsing through results and hence to improve interaction with the user. Section 4 introduces two prototypical user Web interfaces to demonstrate how similarity-based information retrieval can be integrated into a SDI.

4 Integration into SDI

This section presents two conceptual designs for Web user interfaces implementing the similarity-based retrieval paradigms *e* and *h* for the canoeing scenario[10]. An architecture and workflow for the integration of SIM-DL into an SDI is discussed and the requirements for such integration are pointed out.

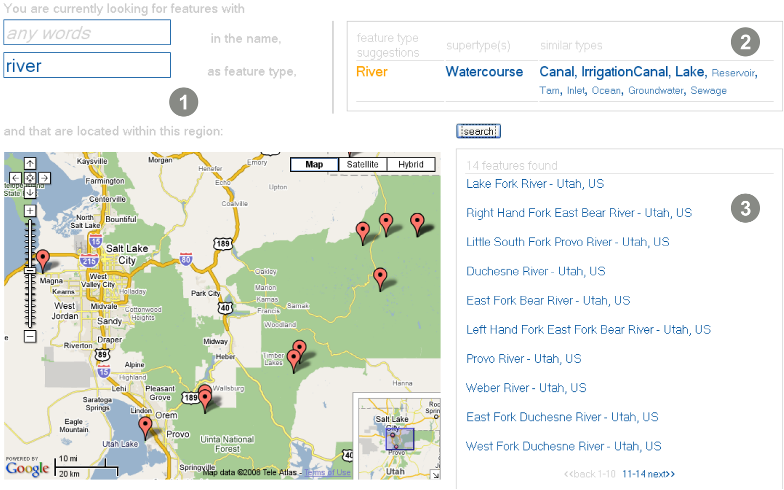
4.1 Similarity-enabled User Interfaces

Figure 4 displays a user interface implementing paradigm *e*. According to the canoeing scenario and using this interface, Nicole searches for features of any name that are of type *River* (C_s) and located near Park City, Utah (*1*).

The context concept C_c is defined as LCS of all feature types which have features in the map extent (and is hence set to $Waterbody$ with respect to our ontology fragment). As result, a tag cloud showing alternative (similar) feature types is returned (*2*). After clicking on the search button, features within the map extent of type *River* are displayed (*3*). The tag cloud can now be used to browse for features that have a different but similar type than $River$.

Figure 5 shows a user interface implementing paradigm *h*. Here, Nicole first specifies known reference features (from Canada) and the map extent (Utah) (*1*). After clicking on the search button, a tag cloud of feature types (C_{t_j}) that are similar to the feature types (C_{r_i}) of the reference features (I_{r_i}) is displayed (*2*). The results are shown for the most similar feature type which is $Canal$, and can be used to display features of other types such as $River$ (*3*).

[10] Note that implementations for both paradigms exist/are under way for different scenarios: paradigm *e* has been implemented within a similarity-aware gazetteer interface [7]; paradigm *f* is currently being implemented for a similarity-aware climbing route recommendation service. See http://sim-dl.sourceforge.net/applications/ for details and source codes.



You are currently looking for features with

any words in the name,

river as feature type,

and that are located within this region:

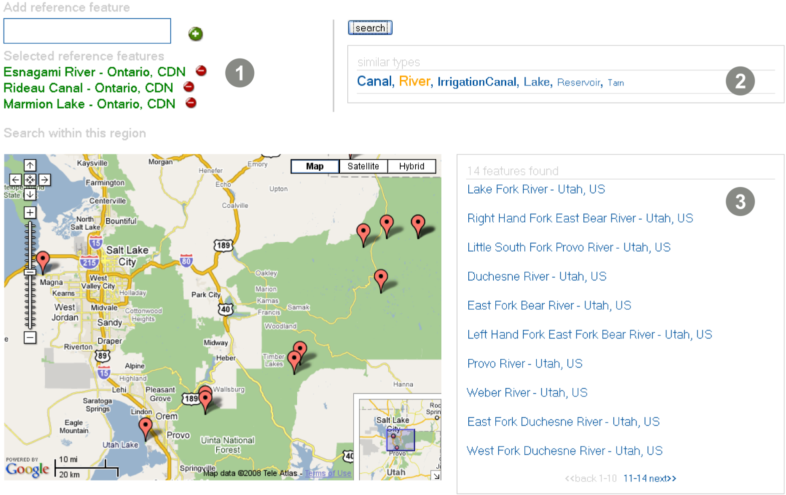
feature type suggestions	supertype(s)	similar types
River	Watercourse	Canal, IrrigationCanal, Lake, Reservoir, Tarn, Inlet, Ocean, Groundwater, Sewage

14 features found

- Lake Fork River - Utah, US
- Right Hand Fork East Bear River - Utah, US
- Little South Fork Provo River - Utah, US
- Duchesne River - Utah, US
- East Fork Bear River - Utah, US
- Left Hand Fork East Fork Bear River - Utah, US
- Provo River - Utah, US
- Weber River - Utah, US
- East Fork Duchesne River - Utah, US
- West Fork Duchesne River - Utah, US

<<back 1-10 11-14 next>>

Fig. 4. A conceptual design of a user Web interface illustrating approach *e*.



Add reference feature

Selected reference features

- Esnagami River - Ontario, CDN
- Rideau Canal - Ontario, CDN
- Marmion Lake - Ontario, CDN

Search within this region

similar types

Canal, River, IrrigationCanal, Lake, Reservoir, Tarn

14 features found

- Lake Fork River - Utah, US
- Right Hand Fork East Bear River - Utah, US
- Little South Fork Provo River - Utah, US
- Duchesne River - Utah, US
- East Fork Bear River - Utah, US
- Left Hand Fork East Fork Bear River - Utah, US
- Provo River - Utah, US
- Weber River - Utah, US
- East Fork Duchesne River - Utah, US
- West Fork Duchesne River - Utah, US

<<back 1-10 11-14 next>>

Fig. 5. A conceptual design for a user Web interface illustrating approach *h*.

4.2 SDI Architecture and Workflow

In figure 6, an SDI architecture for implementing similarity-based information retrieval following paradigms *e*, *f*, or *h* is sketched. It assumes a thematic portal as a client application, e.g., a portal serving hydrology information, with a user interface resembling the ones presented in the previous section. The other components in the architecture are standard WMS and WFS instances as well as a catalogue service including a feature type catalogue (CS-W & FTC) and the WSS. We assume that both services as well as the client have access to and use the same ontology. We do not consider here, how the access to the ontology could be enabled through a service interface (as proposed, e.g., in [32]).

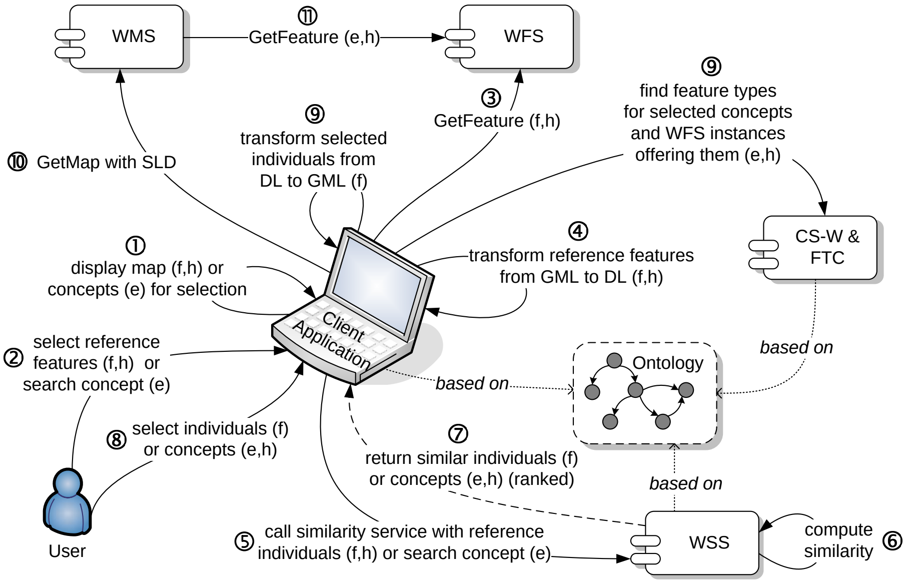


Fig. 6. Architecture and workflow for integrating similarity-based information retrieval into an SDI.

The figure shows the workflow for all three paradigms. Until the similarity is computed, the workflow is the same for approaches *f* and *h*, where reference individuals, rather than reference concepts as in *e*, have to be specified. After this step, approaches *e* and *h* share the same workflow as the similarity computation yields concepts rather than individuals as in approach *f*.

In the first step, the client application displays a map using different feature types from the hydrology domain (*f, h*)[11] or a list of possible search concepts from the ontology (approach *e*). The user can then graphically select one or several features in the map (*f, h*) or a concept from the ontology (*2*). Approaches *f* and *h* require that the selected features are retrieved from the WFS (*3*) and translated from their GML representation into DL individuals (*4*). The DL individuals (*f, h*) or concepts (*e*) that are now available in the client are sent to the WSS (*5*), which computes the similarity using the respective approach (*6*). This yields a ranked list of similar individuals (*f*) or concepts (*e, h*) that is displayed in the

[11] In an SDI architecture, this will be done using a WMS request, possibly using a styled layer descriptor (SLD) document with references to remote WFS instances. These steps are omitted in figure 6 for readability.

client (*7*). The user can then select one or several of the presented individuals (*f*) or one of the presented concepts (*e, h*) (*8*). In approach *f*, the selected individuals are translated back into GML, in approaches *e* and *h*, the CS-W is queried for WFS instances offering feature types annotated with the selected concepts (*9*). Based on the GML or WFS instances, the client builds a SLD document and calls a WMS GetMap operation (*10*). The map is created from the GML directly (*f*) or based on a GetFeature requests to the WFS instances listed in the SLD (*e, h*) (11).

4.3 The Missing Pieces

The architecture and workflow sketched in the previous section puts a number of requirements on the used SDI components, in particular the CS-W and WFS.

Catalogue Service The catalogue service needs to store metadata about three types of resources: (1) services, (2) data, and (3) feature types, as well as the relationships between them. The ebRIM catalogue profile for the CS-W allows storing this information in one registry as well as queries combining them. The ebRIM basic extension package describes the relationship between services and datasets through the *OperatesOn* Association defined in ISO 19119 [33]. In [14], an ebRIM extension package is described that allows the storage of feature catalogue (as defined in ISO 19110 [13]) metadata in an ebRIM catalogue. However, there is no association between the feature types and services and/or data. Such an association would be required in order to find WFS instances that provide a specific feature type.

The model defined in ISO 19110 only includes an optional *definition* attribute of type `string` to describe a feature type. In order to do similarity-based search, a link needs to be established to a concept that annotates the feature type. This link should be stored in the (feature) catalogue rather than the ontology in order to avoid having to update the ontology every time a new feature type is registered in the feature catalogue.

Web Feature Service For the intensional paradigm, the WFS does not have to be changed as the approach works at the feature *type* (rather than the feature *instance*) level and the concepts in the ontology only annotate *features* (rather than also their attributes and/or operations). Thus, once one or several feature types have been discovered, a normal GetFeature request can be sent using the selected feature type (cf. steps 3 and 11 in figure 6).

For the extensional paradigm, features need to be translated into ontology individuals and vice versa (cf. steps 4 and 9). This could be done following the approach described in [34], i.e., by simply mapping the GML properties to DL properties. If the approach is to be successful, this mapping should, wherever possible, use DL roles already existing in the ontology. Otherwise the similarity to existing concepts will be very low. The mapping should preferably be defined by the data provider, or – if no mapping is yet available – by the requester. How to support the creation of such mappings is an open research question.

5 Conclusions and Further Work

This paper investigates paradigms for similarity-based information retrieval, presents prototypical Web user interfaces applying these paradigms, and discusses their integration into SDI as well as remaining difficulties. While the intensional paradigm e has been implemented within SIM-DL and used for a gazetteer research scenario before [7, 35], the integration of the new extensional paradigm f within SIM-DL is under development. Further research should especially focus on approach h. It allows to compute inter-concept similarity without requiring the user to define the search and context concept manually. In many application areas, selecting reference individuals, i.e., examples, may be more intuitive both in terms of using the Web user interface as well as in interpreting the results. In several cases, such as the gazetteer scenario [7, 35] and to a certain degree also the SDI integration discussed here, geographic features are not available as instances within the ontology; hence paradigm h can be used instead of f (and e) to deliver similar feature types. While the question which of the presented search paradigms (also including those purely based on subsumption reasoning) fits best depends on the application area, it would be fruitful to analyze whether certain scenarios abet a particular paradigm. This could be done by human participants tests, but also by classifying the scenarios. For instance, the benefit of similarity lies in browsing through potential results and reducing the complexity of user interfaces [35], while scenarios which require guaranteed results (e.g., in emergency scenarios) may put more focus on subsumption reasoning. Additionally, the list of subsumption and similarity-based paradigms presented in this paper is not exclusive. On may also think of using logical negation to define the search concept. As pointed out by Nedas and Egenhofer [24], the interpretation of such queries is not trivial in terms of a similarity ranking.

Further research should also focus on how to display the retrieved features and types to the user. While SIM-DL supports value rankings, tag clouds, and categories so far [18], other visualization and interaction methods have to be investigated. For instance (for paradigm f), one may think of sliders to select the similarity threshold value above which features should be displaying on the map.

References

1. Lutz, M., Klien, E.: Ontology-based retrieval of geographic information. International Journal of Geographical Information Science **20**(3) (2006) 233–260
2. Stoimenov, L., Djordjevic-Kajan, S.: An architecture for interoperable gis use in a local community environment. Computers & Geosciences **31** (2005) 211–220
3. Möller, R., Haarslev, V., Neumann, B.: Semantics-based information retrieval. In: Proc. IT&KNOWS-98: International Conference on Information Technology and Knowledge Systems, 31. August- 4. September, Vienna, Budapest. (1998) 49–56
4. Küsters, R.: Non-Standard Inferences in Description Logics. Volume 2100 of Lecture Notes in Artificial Intelligence. Springer (2001)

5. d'Amato, C., Fanizzi, N., Esposito, F.: A semantic similarity measure for expressive description logics. In: CILC 2005, Convegno Italiano di Logica Computazionale, Rome, Italy (2005)
6. Janowicz, K.: Sim-dl: Towards a semantic similarity measurement theory for the description logic $\mathcal{ALCNR}$ in geographic information retrieval. In Meersman, R., Tari, Z., Herrero, P., eds.: On the Move to Meaningful Internet Systems 2006: OTM 2006 Workshop SeBGIS. Proceedings. Volume 4278 of Lecture Notes in Computer Science. Springer, Montpellier, France (2006) 1681 – 1692
7. Janowicz, K., Keßler, C., Schwarz, M., Wilkes, M., Panov, I., Espeter, M., Baeumer, B.: Algorithm, Implementation and Application of the SIM-DL Similarity Server. In: Second International Conference on GeoSpatial Semantics (GeoS 2007). Number 4853 in Lecture Notes in Computer Science, Springer (2007) 128–145
8. Araújo, R., Pinto, H.S.: Towards semantics-based ontology similarity. In Shvaiko, P., Euzenat, J., Giunchiglia, F., He, B., eds.: Proceedings of the Workshop on Ontology Matching (OM2007) at ISWC/ASWC2007, Busan, South Korea. (2007)
9. Nebert, D.D.: Developing spatial data infrastructures: The SDI cookbook (2004) Available from http://www.gsdi.org/docs2004/Cookbook/cookbookV2.0.pdf.
10. McKee, L. In: Who wants a GDI? Oxford University Press (2000) 13–24
11. Bernard, L., Kanellopoulos, I., Annoni, A., Smits, P.: The european geoportal - one step towards the establishment of a european spatial data infrastructure. Computers, Environment and Urban Systems **29** (2005) 15–31
12. European Commission: Spatial data infrastructures in europe, state of play spring 2005: Summary report of activity 5 of a study commissioned by the ec (eurostat & dgenv) in the framework of the inspire initiative. Technical report, European Commission, Brussels (2005)
13. ISO: ISO 19110:2005 geographic information – methodology for feature cataloguing. International standard, ISO TC 211 (2005)
14. OGC: Feature type catalogue extension package for ebRIM (ISO/TS 15000-3) profile of CSW 2.0. Discussion Paper OGC 07 172r1, Open Geospatial Consortium Inc. (2007)
15. Goldstone, R.L., Son, J.: Similarity. In Holyoak, K., Morrison, R., eds.: Cambridge Handbook of Thinking and Reasoning. Cambridge University Press (2005)
16. Medin, D., Goldstone, R., Gentner, D.: Respects for similarity. Psychological Review **100**(2) (1993) 254–278
17. Tversky, A.: Features of similarity. Psychological Review **84**(4) (1977) 327–352
18. Janowicz, K.: Kinds of contexts and their impact on semantic similarity measurement. In: 5th IEEE Workshop on Context Modeling and Reasoning (CoMoRea'08) at the 6th IEEE International Conference on Pervasive Computing and Communication (PerCom'08), Hong Kong, IEEE Computer Society (March 2008)
19. Keßler, C., Raubal, M., Janowicz, K.: The effect of context on semantic similarity measurement. In Meersman, R., Tari, Z., Herrero, P., eds.: On the Move to Meaningful Internet Systems 2007: OTM 2007 Workshop SWWS 2007. Number 4806 in Lecture Notes in Computer Science, Vilamoura, Portuga, Springer (2007) 1274–1284
20. Frank, A.U.: Similarity measures for semantics: What is observed? In: COSIT'07 Workshop on Semantic Similarity Measurement and Geospatial Applications, Melbourne, Australia (2007)
21. Rodríguez, A., Egenhofer, M.: Comparing geospatial entity classes: an asymmetric and context-dependent similarity measure. International Journal of Geographical Information Science **18**(3) (2004) 229–256

22. Raubal, M.: Formalizing conceptual spaces. In Varzi, A., Vieu, L., eds.: Formal Ontology in Information Systems, Proceedings of the Third International Conference (FOIS 2004). Volume 114 of Frontiers in Artificial Intelligence and Applications. IOS Press, Amsterdam, NL (2004) 153–164
23. Li, B., Fonseca, F.: Tdd - a comprehensive model for qualitative spatial similarity assessment. Spatial Cognition and Computation **6**(1) (2006) 31–62
24. Nedas, K., Egenhofer, M.: Spatial similarity queries with logical operators. In Hadzilacos, T., Manolopoulos, Y., Roddick, J., Theodoridis, Y., eds.: SSTD '03 - Eighth International Symposium on Spatial and Temporal Databases, Santorini, Greece. Volume 2750 of Lecture Notes in Computer Science. (2003) 430–448
25. Gahegan, M., Agrawal, R., Banchuen, T., DiBiase, D.: Building rich, semantic descriptions of learning activities to facilitate reuse in digital libraries. International Journal on Digital Libraries **7**(1) (2007) 81–97
26. Horrocks, I.: Implementation and optimization techniques. In: The description logic handbook: theory, implementation, and applications. Cambridge University Press, New York, NY, USA (2003) 306–346
27. Wilkes, M., Janowicz, K.: A brief intro to the sim-dl dig extension. Technical report, Institute for Geoinformatics, University of Münster, Germany (2008)
28. Barsalou, L.: Ad hoc categories. Memory & Cognition **11** (1983) 211–227
29. Barsalou, L.: Intraconcept similarity and its implications for interconcept similarity. In: Similarity and analogical reasoning. Cambridge University Press (1989) 76–121
30. Janowicz, K., Keßler, C., Panov, I., Wilkes, M., Espeter, M., Schwarz, M.: A study on the cognitive plausibility of SIM-DL similarity rankings for geographic feature types. In Bernard, L., Friis-Christensen, A., Pundt, H., eds.: 11th AGILE International Conference on Geographic Information Science (AGILE 2008). Lecture Notes in Geoinformation and Cartography, Girona, Spain, Springer (5-8. May 2008) 115–133
31. Cross, V., Sudkamp, T.: Similarity and Computability in Fuzzy Set Theory: Assessments and Applications. Volume 93 of Studies in Fuzziness and Soft Computing. Physica-Verlag (2002)
32. Lacasta, J., Nogueras-Iso, J., Béjar, R., Muro-Medrano, P.R., Zarazaga-Soria, F.J.: A web ontology service to facilitate interoperability within a spatial data infrastructure: Applicability to discovery. Data & Knowledge Engineering **63**(3) (2007) 945–969
33. ISO: ISO 19119:2005 geographic information - services. International standard, ISO TC 211 (2005)
34. Klien, E.: A rule-based strategy for the semantic annotation of geodata. Transactions in GIS, Special Issue on the Geospatial Semantic Web **11**(3) (2007) 437–452
35. Janowicz, K., Keßler, C.: The role of ontology in improving gazetteer interaction. International Journal of Geographical Information Science (IJGIS) **10** (2008; forthcoming)