Using Semantic Similarity to Improve Information Discovery in Spatial Data Infrastructures

Fabio G. de Andrade, Cláudio de S. Baptista

[1] Instituto Federal de Educação, Ciência e Tecnologia da Paraíba, Brazil
fabio@ifpb.edu.br
[2] Universidade Federal de Campina Grande,Brazil
baptista@dsc.ufcg.edu.br

Abstract. In the recent years, several works have been proposed with an approach to the use of semantics to improve the process of discovering geographic resources offered by spatial data infrastructures. However, semantic queries may return a large number of results, what causes the necessity for efficient ways to evaluate the relevance of each retrieved result. This article proposes a framework that uses ontologies and thematic relevance to suggest a measurement that allows evaluating how relevant is each resource offered by the infrastructure to the user's query. This feature allows the results retrieved in a query to be organized through a ranking, in such a way that the most relevant resources are presented to the user first.

Categories and Subject Descriptors: H. Information Systems [**H.3 Information Storage and Retrieval**]: H.3.3 Information Search and Retrieval

Keywords: semantic, similarity, ontology, spatial data retrieval, spatial data infrastructures

1. INTRODUCTION

In recent years, spatial data infrastructures (SDIs) [Williamson et al. 2003] have been developed in order to ease the discovery and interoperability of geographic data supplied by different information sources. The development of open standards to the geospatial domain has reduced the problems concerning interoperability. However, discovering geographic data which are already available is still a hard task.

A limitation of current infrastructures is that their catalog services perform queries based only on keywords. This can lead to the execution of queries with low recall, as the resources described in terms related to the keywords used to generate the query itself are not retrieved. Also, low precision is obtained, since many irrelevant resources which have the same descriptive terms as those of the query end up by being retrieved. One way to overcome this limitation is to apply semantic web concepts to improve discovery of geospatial data. The objective of the semantic web [Berners-Lee et al. 2001] is to use formal means to describe the semantics of resources published in the web, improve information discovery and promote data sharing among applications. The semantic web principles have been implemented through ontologies [Guarino 1995], which are formal conceptualizations of an application domain. Ontologies are most advantageous because they make the semantics of an application domain understandable to both human and machines. In recent years, several works have been developed by using semantic web concepts to improve discovery and integration of geographic data. The application of semantic web ideas to the geospatial domain has been termed geospatial semantic web[Kuhn 2005][Egenhofer 2002].

Copyright©2011 Permission to copy without fee all or part of the material printed in JIDM is granted provided that the copies are not made or distributed for commercial advantage, and that notice is given that copying is by permission of the Sociedade Brasileira de Computação.

Usually, applications that use ontologies to discover information adopt an approach based on a semantic relationship known as subsumption. This kind of solution consists in locating all resources related to concepts that are subsumed by the search concept defined in the user's query. The biggest advantage of this kind of solution is that it improves the performance of queries, since search engines may understand the semantic relationships that exist between the search concept and the concepts used to annotate the resources that are offered by a provider. Nevertheless, this kind of solution has an important limitation, since it considers that all of the retrieved results have the same relevance to the user. For example, if a user requests feature types about a specific concept, feature types associated to subclasses of this concept (which offer only a part of the information requested) are considered as relevant as the feature type annotated exactly with the concept, which is probably more important to the user. Thus, during the presentation of the query result, more relevant resources can be showed later to the user. In the worst case, since queries can retrieve a large number of results, these resources can be presented too late and may not even be evaluated by the user. This feature produces the necessity for the development of mechanisms that permit a search machine to evaluate the relevance of each retrieved resource.

To tackle this limitation, this article proposes an approach that uses similarity to improve resource discovery in SDIs. The main contribution consists of the development of a measurement that enables to evaluate the relevance of each feature type offered by the SDI to the user's query. Furthermore, it proposes a new measurement to evaluate similarity among concepts defined in ontologies, and shows how some ideas of classic information retrieval can be reused and adapted to the geospatial domain.

The remainder of the article is organized as follows. Section 2 discusses related work. Section 3 addresses information discovery in spatial data infrastructures. Section 4 describes the approach used to evaluate the relevance of resources. Section 5 highlights the information discovery process. Section 6 addresses implementation issues and the results obtained. Finally, section 7 concludes the article and highlights further work to be undertaken.

2. RELATED WORKS

Over the years, several studies have been proposed to implement the discovery of information in both SDIs and geographic portals. Although these works are related to the same research area, they differ from one another with respect to the type of resources discovered and the type of approaches used to retrieve these resources.

Some studies to help the discovery of geographic services have been proposed. Klien et al. [Klien et al. 2006] recommend a solution that allows the discovery of services used to improve the management of disasters. Their approach employs descriptors based on XML, where the output parameters of a service are linked to concepts defined in ontologies. Information discovery is obtained by selecting the services which contain, in their descriptions, the search concept required by the user, or a concept subsumed by this concept. In another work [Stock et al. 2010], a feature type catalog was used to describe a set of feature types and the semantic relationships that exist among them. Furthermore, the catalog defines a set of operations which can be performed for each feature type. This catalog holds a link to the services that implement these operations. The work offers a set of services enabling users to find the concepts defined by the catalog, select an operation defined for a type and retrieve a list of the services that implement that operation. On the other hand, Lutz [Lutz 2007] has developed a solution according to which services are annotated by means of semantic signatures described by the use of first-order logic. Such signatures use ontologies to describe several features of the service, such as input and output parameters, the conditions to be satisfied as to its execution, and the effects derived from its invocation. Lemmens et al. [Lemmens et al. 2007] have proposed an ontology that describes a hierarchy of geographic services. When new services are registered, these are related to the class that represents their functionality; moreover, an OWL-S [Martin et al. 2004] profile is used to describe its properties. During the searching process, the service ontology is used to find

services or service compositions to implement the functionality requested by the user. After that, the search engine locates the service implementations that will implement the services selected during the searching process. Semantics is used in all instances in order to improve resource discovery in SDIs. However, they focus entirely on service discovery; they do not deal with feature types discovery.

As to feature types discovery, some solutions have been developed to it. Janowicz et al. [Janowicz et al. 2008] propose a similarity-based solution to discover feature types supplied by SDIs. Their work is based on a framework [Janowicz 2006] that evaluates the similarity between the concepts of ontologies described in Description Logic [Baader et al. 2003]. In this work, the feature types offered by services are associated to the concepts defined in domain ontologies. During the searching process, the filters select the concepts that are most similar to the concept defined by the query. Then, the user can select one of these concepts to obtain the features associated to it. The biggest advantage of this work is that the values of similarities between concepts can be used to determine the relevance of each feature type. However, the work is directed towards a very specific ontology. Besides, it takes into account the fact that resources annotated with the same concept are evaluated with the same degree of relevance. This represents a drawback, since each concept may be related to a higher quantity of resources. Another approach, based on similarity, was proposed by Zhang et al. [Zhang et al. 2010]. In that work, the importance of a feature type being evaluated in a user's query is determined by the sum of the similarity degree of the following characteristics: the ontology concepts that they are related to, their properties, geometries (such as point, line and polygon), and bounding-boxes. Nevertheless, the work takes into account only concepts related through generalization/relationship so as to determine semantic similarities. Lutz and Klien [Lutz and Klien 2006], on the other hand, propose a solution where each feature type and its respective properties can be associated to concepts defined in ontologies along mapping records [Bowers and Ludäscher 2004]. In this work, information discovery is achieved via subsumption without the necessity of metrics to rank the resources discovered.

Lutz and Kolas [Lutz and Kolas 2007] developed a method that employs rules to perform the semantic annotation and the retrieval of spatial data spread over various data sources. Their approach uses two kinds of rules. The first is used to define the data model of each information source; whereas the second is used to describe how the data model adopted by the data source may be mapped on to a global ontology that is shared by all data sources in the infrastructure. To perform information discovery, their approach retrieves, from each data source, the relevant data to the query. All the retrieved data are loaded on a knowledge base, and reasoning is applied to this base in order to infer the data which solves the user's query. Smits and Friis-Christensen [Smits and Friis-Christensen 2007] developed a method for the discovery of data in SDIs. In their work, the semantic annotation of the resources is done by associating these resources to concepts defined in a thesaurus. The user can then browse the terms of the thesaurus to discover correlated resources. Nevertheless, the work does not use a ranking to evaluate the relevance of each retrieved resource. Other important works have the same limitation [Athanasis et al. 2009], [Lutz et al. 2008] and [Wiegand and Garcia 2007].

The analysis of related work shows that information discovery in SDIs is still an open problem. The use of ontologies allows exploring semantics to enhance the quality of query results. However, queries can produce a large number of results which need to be evaluated by the user before being completely retrieved. This problem leads to the necessity of developing efficient mechanisms to evaluate the relevance of each retrieved result. Such a solution is based on a ranking approach, in which the most relevant resources are presented first. This ranking reduces the time spent during the result evaluation process and facilitates both data discovery and reuse.

3. INFORMATION DISCOVERY IN SPATIAL DATA INFRASTRUCTURES

Before describing the proposed approach for evaluating the relevance of spatial resources supplied by SDIs, it is necessary to understand which kind of information can be discovered. Aiming to standardize the access to geographic data offered by different data sources, SDIs usually publicize their dataset

through a set of standardized web services. Those services are specified by the Open Geospatial Consortium (OGC), which is an organization that develops standards for the geospatial domain, and whose interfaces are public and open. Important examples of OGC web services to improve the retrieval of geographic data are the Web Map Service (WMS)[1], which allows access to vector map layers; the Web Feature Service (WFS)[2] to permit access to spatial data in GML format; and the Web Coverage Service(WCS)[3] to enable access to coverage in raster format. Since standardization is a key issue in the development of spatial data infrastructures, the use of OGC web services as a way to give access to spatial data has become more popular in recent years. Furthermore, the development of tools to facilitate the offering of geographic data by means of such services contributes enormously to their popularization. For example, the federal SDI of the United States offers more than 13,000 services, mostly WMS and WFS. This feature permits us to think of SDIs as a collection of geographic services. For the sake of simplicity, we use the term OGC services when referring to the OGC services used to access spatial data, even knowing of the existence of OGC services which are directed at other purposes.

An important feature of the OGC services is that they give access to a set of spatial data that is offered by a data provider. Each geographic data type offered by a dataset is called feature type, and offers a set of georeferenced data about a certain theme. For example, the service offered by an agency specialized in water resources may provide feature types about rivers, water bodies and watersheds; whereas a service offered by a disaster management organization can render feature types such as fires, tsunamis and flooding. These features make it necessary to develop a mechanism that enables service users to get information about the feature types supplied by the service. In order to attend to this requirement, OGC services offer an operation called getCapabilities. When invoked, this operation produces a document describing the general features of the service as well as a description of each one of its feature types. For each feature type, the document returned by the service contains information such as name, textual description, bounding-box, keywords and the coordinate reference system. Each service also offers operations that enable users to access the spatial data offered by a feature type. Nevertheless, the operation that must be used to retrieve spatial data depends on the type of service that is being used.

The use of OGC services improves accessibility to spatial data, but they do not solve the problem of finding data. To become a data discoverable service, providers must register their services in the SDI. During this process, the provider should give information to describe both the service and the feature types offered by it. Because of standardization, the information to be informed during registry is that contained in the standard metadata adopted by the infrastructure. Each infrastructure is free to develop its metadata standard. However, the patterns adopted are usually reduced to a profile of either ISO 19115[4] or FGDC[5].

Once the metadata provided by service providers are stored, they must be made available to enable users to find the information of their interest. This task is carried out by means of a catalog service. The catalog service[6] enables users to locate the resources offered by an infrastructure. When a user poses a query to this service, the metadata registered in the SDI is promptly evaluated and returned, after which all the metadata records that match the constraints are defined in the query. After this the user may evaluate the results to identify which of the returned resources fits his/her necessities better. Although the use of catalog services has improved information discovery in SDIs, it does not solve the problem completely. Some of the remaining problems are:

[1]OGC WMS: http://portal.opengeospatial.org/files/?artifactid=4756
[2]OGC WFS: https://portal.opengeospatial.org/files/?artifactid=8339
[3]OGC WCS: http://portal.opengeospatial.org/files/?artifactid=27297
[4]ISO Geographic Information Metadata: http://www.iso.org/iso/isocatalogue/cataloguetc/cataloguedetail.htmcsnumber=26
[5]FGDC: http://www.fgdc.gov/standards/projects/FGDC-standards-projects/metadata/basemetadata/v20698.pdf
[6]OpenGIS Catalogue Services Specification: http://portal.opengeospatial.org/files/?artifactid=21460

—**Loss of information:** when service providers register their services in the SDI, they usually provide metadata that describe the service as a whole, without supplying detailed information about each feature type the service is composed of. Furthermore, since many services provide a large quantity of feature types, many providers give information about part of the feature types offered by the service. Since information discovery is performed using the information contained in the metadata, many feature types cannot be retrieved. This is because their descriptions have not been supplied by their providers during the registry process;

—**Discovery at service level:** currently, when the user poses a query to the catalog service, the service generally returns the metadata on the services that match query restrictions, after which the user will choose a service and invoke its getCapabilities operation to discover details from among its feature types, the ones that offer the desired information. This process can be quite tedious, since a query can return a large quantity of services and each service can provide a large quantity of feature types; and

—**Imprecise information:** some information passed on by service providers during the registry process describes the dataset as a whole, and do not offer detailed information about each feature type. For example, the service metadata record contains a bounding-box value that describes the spatial extension of its dataset. Nevertheless, more precise information about the geographic extension of a feature type can be retrieved in the document returned by the getCapabilites operation. Since information discovery is obtained from the information contained in the metadata, the user can invoke this operation, analyze the retrieved document and verify which returned feature type covers the geographic region in which he/she is interested. The same problem occurs with other attributes, such as the temporal extension.

To overcome the limitations discussed above, we propose a model that permits users to easily locate the feature types offered by the SDI. Furthermore, our approach takes advantage of the way these types are currently offered by services to estimate a relevance value for each feature type retrieved by a query.

4. AN APPROACH TO EVALUATE FEATURE TYPE RELEVANCE

The main contribution of the present work is the description of a similarity metric that permits us to verify how similar are the concept used to annotate the data type under evaluation and the search concept. The approach proposed is implemented in a prototype created to evaluate its performance. The verification is carried out in three stages:

—to verify how much the concept used to annotate the data type which is under evaluation is similar to the search concept defined by the user;

—to verify how relevant the theme requested by the query is to the spatial service that offers the data type under evaluation; and

—to combine the values of both measurements to evaluate the relevance of the data type under evaluation for the query.

The next subsections describe each one of these stages. Nevertheless, we will proceed with an overview of how the information used during the searching process is collected.

4.1 Gathering information

The approach proposed by this article focuses on the discovery of the feature types offered by WMS and WFS services, in which a feature type can either be a vector map layer or a GML feature type, depending on the kind of service. To reach this objective, we must gather information about all feature types provided by each service registered at the infrastructure. This process is done when

the data source is registering its resources in the infrastructure. When this occurs, we automatically invoke the getCapabilities operation, in order to retrieve a XML document containing a list of all the feature types supplied by the service. After parsing this document, we collect as much information as possible about each feature type, such as: name, title, textual description, the type (vector map layer or feature type) and the bounding-box of the geographic region it covers.

After the information is retrieved, the prototype that implements our solution shows the user that is registering the service a page showing all feature types available. Then, the user is asked to perform the semantic annotation of each one of these types. For that, the user must choose, from among the existing concepts in the domain ontologies used by the infrastructure, the ones that better represent the information supplied by that feature type. For example, a feature type which offers information about water reservoirs can be annotated under the concept *WaterCourse*, while one which describes only rivers can be annotated under the concept *River*. Once all feature types are annotated, the registering is finished, and the information concerning the service and its feature types are permanently stored in a relational database.

4.2 Evaluating the similarity between the concepts

The first stage of the process used to measure the relevance consists in evaluating the similarity between the search concept of the user's query and the concept chosen for the semantic annotation of the data type under analysis. The approach used to carry out this task has the following characteristics:

—**support to other types of relationships:** many works that propose to evaluate the similarity between concepts consider just the generalization relationship. However, other relationship types, such as composition, cannot be neglected, since they denote an association between the concepts involved. As the composition is not such a strong relationship as the generalization, weights are necessary to distinguish the relationship types. For example, two concepts associated by a generalization relationship must have a degree of similarity greater than that existing between two concepts associated by the composition relationship;

—**asymmetry:** symmetrical similarity measurements consider that the similarity between two pairs of concepts is the same, independently of the comparison order. However, for the problem studied in this article, we considered that the symmetry is not an interesting feature. For example, let us suppose that a concept B is a sub-concept of a search concept A. We may state that all data associated to B are relevant to the user, since every instance of concept B is also an instance of concept A. Nevertheless, if the user is looking for data associated to concept B, not all data associated to concept A are relevant to the user, since not all instances of concept A are instances of concept B. This characteristic requires, in the second case, that the symmetry must be smaller than in the first case, due to the existence of information of no interest to the user. The same idea is applied to the composition relationship; and

—**degree of generalization:** ontologies are described through concepts that are organized in a hierarchical form, through generalization relationships. Let us suppose that there is a hierarchy from a search concept (SC) in the ontology under analysis. As we go through this hierarchy, we find concepts that are more and more specialized with respect to SC and, consequently, have more difference with respect to it. Thus, the similarity of concepts must decrease gradually as the deepness of the concept increases.

The evaluation of the similarity between concepts is performed through a semantic network, generated from the parsing of the ontology at the time it is added to the SDI. The construction of this network takes into consideration two types of semantic relationship existing between the ontology concepts: inheritance and composition. The following algorithms in Figure 1 present how a semantic network may be generated. The first algorithm is used to start the process of generating the network and the second one to expand the production of the network to new concepts obtained from new

concepts which are processed by the algorithm. In the first algorithm, the network is generated from each root concept in the ontology. A root concept is a concept that has no superclass in the ontology.

The result of the execution of the algorithms above is a semantic network which contains all of the concepts defined in the ontology and the semantic relationships existing between these concepts. In such network, nodes represents concepts, and links represent semantic relationships. The network produced has two kinds of link: one to define inheritance relationships and other to define composition relationships. Figure 2 shows a semantic network produced for the hydrographic domain, extracted from the GEMET ontology.

After the semantic network is produced, the framework evaluates the degree of similarity for all combinations of pairs of concepts defined in the ontology. This similarity is calculated through the analysis of the path that connects the two concepts under evaluation in the semantic network. The calculation of this similarity is performed taking into consideration two kinds of variables: the semantic relationship between the concepts and the distance between them in the network.

The objective of the semantic relationship between the concepts is to assign a greater degree of similarity to pairs of concepts that have stronger semantic relationships. As inheritance is a semantic relationship stronger than composition, concepts associated by an inheritance relationship must have a degree of similarity greater than that concepts associated by a composition relationship. To implement this constraint, a weight is assigned to each link of the network. For each node, two weights are possible: a normal weight and an inverse weight. The weight used to evaluate the similarity depends on the order of the concepts involved. This constraint is used to keep the asymmetry requirement. The use of two kinds of weight ensures asymmetry, keeping the discovery of paths in graphs simple. Currently, the weight used for normal and inverse weights are, respectively, 0.8 and 0.6, for inheritance relationships, and 0.6 and 0.4, for composition relationships.

To perform the comparison between the two concepts, the first step consists in locating, in the network, the path that connects the two concepts. To allow the comparison between concepts Q and D, there must be at least one path from node Q that leads to D, or vice-versa. When none of these paths exist, the concepts are considered disjoint and the degree of similarity between them is assumed to be zero. When any of these paths can be found, the framework uses weights of the nodes in this path to evaluate the semantic relationship between them. Let $W = \{w_1, w_2, ..., w_n\}$ be the set of the weights of each link in the shortest path that connects the concepts Q and D in the semantic network that represents the ontology in which these concepts were defined. The value of the semantic relationship can be formally defined in Equation 1. In order to ensure asymmetry property, the weight values depend on the order of the concepts in the path. If the path starts with the search concept Q and ends with the concept D, the normal weights are considered. However, if the path starts with the concept D and ends with the concept Q, the inverse weight is considered.

$$semanticRelationship = min\{w_1, w_2, ..., w_n\} \tag{1}$$

The second variable used to compare the similarity between the concepts is the distance between them. Rada et al. [Rada et al. 1989] introduce the semantic distance as a metric to evaluate similarity among concepts in semantic networks. The objective of this variable is to guarantee that pairs of concepts which are closer in the network have a greater degree of similarity compared to more distant pairs. We use this metric to implement the constraint to the degree of generalization between the concepts presented in Section 4.2. This measurement is inversely proportional to the degree of similarity, that is, as the distance between concepts increases, the similarity between them diminishes.

After evaluating the semantic relationship and the distance between the two compared concepts, the values of these variables are used to measure the similarity between the concepts. To calculate these values, a weight is assigned to each of these variables, where w_1 represents the weight assigned

```
generateSemanticNetwork(O: Ontology): SemanticNetwork;
begin
      SN = new SemanticNetwork();
      rootNode = createNode("Thing");
      SN.addNode(rootNode);
      for each RCi in O do
      begin
           newNode = createNode(RCi)
           SN.addNode(newNode);
           SN.addSubclassEdge(rootNode, newNode);
           expandSemanticNetwork(SN, newNode, O);
      end;
      return SN;
end;

expandSemanticNetwork(sn:SemanticNetwork,currentNode:Node, O:ontology):
void;

begin
      SC = O.getSubClasses (currentNode.getConcept());
      for each SCi in SC do
      begin
          newNode = createNode(SCi);
          sn.addNode (newNode);
          sn.addSubclassEdge(currentNode, newNode);
      end;
      OP = O.getObjectProperties(currentNode.getConcept());
      for each OPi in OP do
      begin
          newNode = createNode(OPi.getRange());
          sn.addNode (newNode);
          sn.addAssociationEdge(currentNode, newNode);
      end;
end;
```

Fig. 1. Semantic network generation algorithm

Table I. Similarity matrix

	Hydrosphere(D)	WaterCourse (D)	River(D)	RiverBed (D)
Hydrosphere (S)	1	0.84	0.74	0.54
WaterCourse (S)	0.68	1	0.84	0.58
River (S)	0.58	0.68	1	0.68
RiverBed (S)	0.38	0.42	0.52	1

to the semantic relationship and w_2 represents the weight of distance. The use of these weights makes the similarity between the concepts to be evaluated through the Equation 2.

$$similarity(Q, D) = w_1 * semanticRelationship(Q, D) + w_2 * \frac{1}{distance(Q, D)} \quad (2)$$

The solution evaluates the degree of similarity between all pairs of concepts defined in the ontology (in both directions), generating a similarity matrix. The values of these similarities are stored in a relational database. Table I shows the similarity matrix for an excerpt of the concepts of the semantic network depicted in Figure 2. Concepts marked with S represent the concept defined in the user query, whereas concepts marked with D represent the ones used to annotate the feature type that is being evaluated.

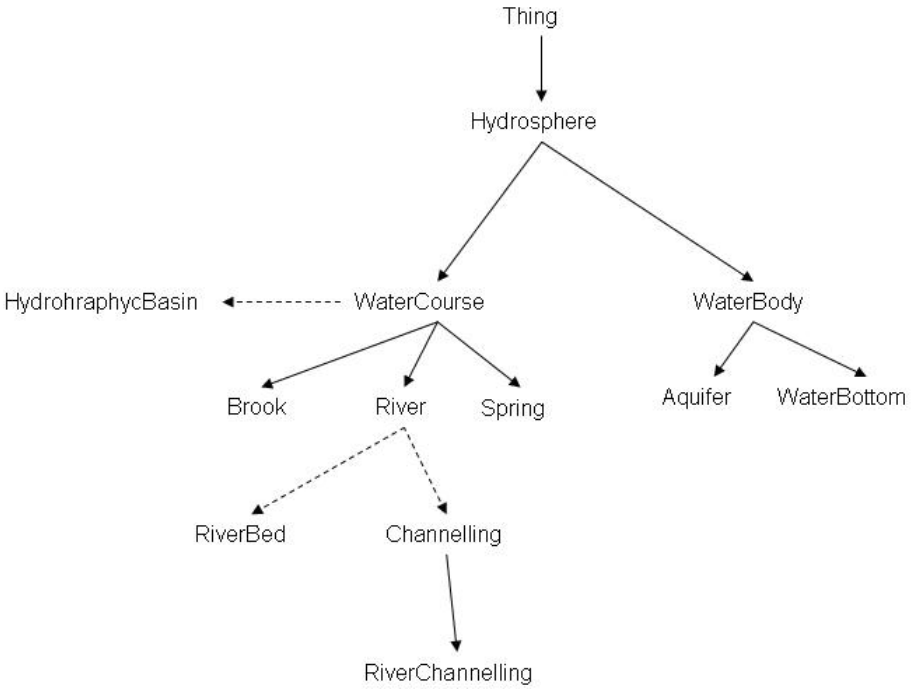

Fig. 2. Semantic network produced for a hydrographic ontology

4.3 Evaluating the degree of thematic relevance

Besides the degree of semantic similarity among the concepts involved in the query, we consider the degree of thematic relevance to improve the discovery process. The objective of this measurement is to evaluate how relevant is a theme requested in a query to the service which offers the data type under analysis. Through this measurement, data types offered by services in which the theme has more relevance are shown first to the user during the presentation of results. The value of this measurement is very important, since many data types are offered by several services, especially if the user's query requests a very general theme.

The degree of relevance that a certain theme has to a service is calculated through the normalized frequency, which is a measurement used in classic information retrieval [Baeza-Yates and Ribeiro-Neto 1999] to evaluate the relevance of a certain term in a document. To evaluate this degree, we compute, at the time the service is registered, the normalized frequency of each theme offered by it. This way, the degree of relevance (tf) of a theme C to a service is calculated through the proportion of the number of times the theme occurs in the service (fi) and the number of data types offered by the service (N). The value of fi for a certain concept C is calculated based on the semantic relationships defined in the ontology. Such calculation comes from Equation 3. In this equation, $fi(C)$ is the number of occurrences of the concept under evaluation, and $fi(S)$ and $fi(SC)$ represent, respectively, the number of occurrences of a synonym concept of C and the number of occurrences of a concept that is a sub-class of C.

$$tf(C,S) = \frac{fi(C) + \sum fi(S) + \sum fi(SC)}{N} \tag{3}$$

Besides the normalized frequency, our model also evaluates the importance of the theme requested by a query to the infrastructure as a whole. This evaluation is done through a metric called inverse

frequency (idf). As in classic information retrieval, the value of this metric is determined from the number of services in the SDI where this theme occurs. In our approach, the inverse frequency of a theme C to the infrastructure is calculated through Equation 4, where N represents the number of service registered in the SDI and ni represents the number of services where the theme C occurs.

$$idf(C) = log\frac{N}{n_i} \tag{4}$$

Once calculated, the values of normalized frequency and inverse frequency are used to determinate the thematic relevance of a feature type T for the user's query. This relevance is calculated through the product of the values for these frequencies. Thus, let us consider C as an ontology concept and S as the service that offers T, the thematic relevance of T is calculated through Equation 5.

$$thematicRelevance(T, C) = tf(C, S) * idf(C) \tag{5}$$

The evaluation of thematic relevance of a feature type is done based on the frequency of the requested theme in each service registered in the infrastructure. To improve the performance of our queries, these frequencies are computed at the moment of service registering. The information related to the number of occurrences of each theme inside the service is stored in the database. Such characteristics allow accelerating query resolution, keeping the solution scalable even when there is a large number of services.

4.4 Calculating the relevance

After defining the metrics used to calculate the degree of relevance of a data type to a user's query, the next step consists in defining how the values of these metrics will be used for that purpose. One possibility to solve this problem would be the representation of the user's query and the data type under evaluation as vectors in a bi-dimensional space and use the Euclidian distance to evaluate the similarity. However, this kind of metric represents similarity through a real number, corresponding to the distance between the points. Thus, in order to solve the problem, we adopted the sum of the values of the metrics, where a weight is assigned to each of the metrics. This technique, besides offering flexibility, since all weights may be altered to perform new queries, also offers similarity values between 0 and 1, which makes the evaluation of similarity more intuitive for the human being.

Thus, given a theme Q defined in the user's query and the theme D associated to the feature type T under evaluation, the degree of relevance of this type for the query is calculated through Equation 6. In such equation, $similarity$ represents the degree of similarity of concepts Q and D, whereas $thematicRelevance$ represents the degree of relevance that the theme D has to the service by which the feature type under evaluation is offered. Finally, w_1 and w_2 represent the weights for each type of measurement in the calculation of semantic similarity. Each weight must have a value between 0 and 1, and their sum must always be equal to 1. Curently, we use the values 0.9 and 0.1, respectively, for w_1 and w_2. Both of these values have been obtained by experiments.

$$semanticSimilarity(Q, T) = w_1 * similarity(Q, D) * w_2 * thematicRelevance(Q, S) \tag{6}$$

5. THE INFORMATION DISCOVERY PROCESS

The model described in the previous section is used as an instrument for ranking feature types during the searching process. To pose a query, the user must define the values for two input parameters: the information type and the geographic extension. The information type represents the kind of

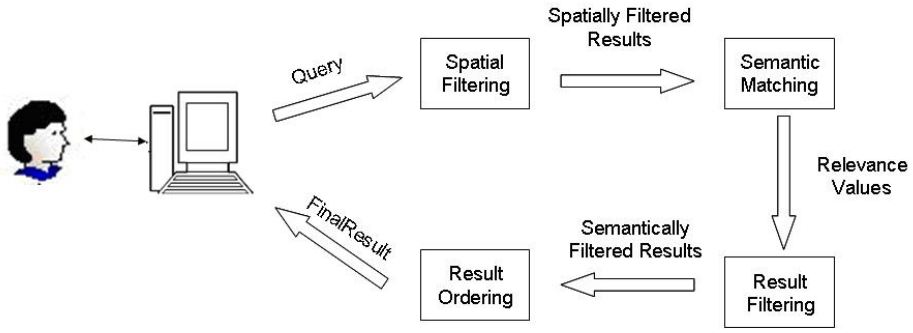

Fig. 3. The information discovery process

information that the user is looking for. In order to define this parameter value, the user must select one of the concepts defined in the ontologies used for the infrastructure. The URI of the selected concept is used as a value for this parameter. The other parameter corresponds to the geographic region which represents the geographic extension that must be covered by the feature type. This information is configured in the form of a bounding-box, which represents the smallest possible rectangle that covers the entire geographic region selected by the user. In case the user has not selected a region, our prototype may use the bounding-box from Brazil as the default value. This option occurs because the Brazilian SDI is being used as a case study to validate the model proposed by this article.

After both the search concept and the geographic extension are defined by the user, the query is forwarded to the search engine. The query is then processed along various stages, as shown in Figure 3. The first stage is that of spatial filtering. The purpose of this stage is to filter, among all the feature types offered by the services registered in the infrastructure, only those whose bounding-box intersects the geographic region defined in the query. This task is carried out by a spatial query executed on the table that stores information about the feature types. In order to speed-up this task, we created our database model through a DBMS with support to spatial data. This kind of DBMS offers the appropriate structures for indexing georeferenced data, such as R-Trees. These indexes enable the search engine to execute spatial queries in very short time intervals.

Once the spatial filtering is done, a list containing all the feature types that intersect the geographic region defined by the query is obtained. In the following stage, we do the semantic matching of these feature types. For this, a matchmaker evaluates the relevance of each feature type according to the equations presented in the previous section. A list containing information about the feature types retrieved in the first stage together with their corresponding relevance values is then produced.

In the third stage, the results obtained from the semantic matching are filtered again. The goal of this stage is to remove feature types irrelevant to the query. To carry out this task, the filter evaluates the relevance value obtained for all resources. Feature types whose ranking value is less than a pre-defined threshold are filtered off the final result. Currently, the value for this threshold is set to 0.2. After the filtering operation, the remaining feature types are sorted out in decreasing order, according to their relevance values. The list obtained at this stage is returned to the search engine, and presented to the user.

6. IMPLEMENTATION AND RESULTS

To evaluate the proposed approach, a prototype was developed. The first step in this implementation was to define the domain ontologies that would be used for semantic annotation and discovery. In our experiments we have used ten domain ontologies, which were created from data models according to

the Brazilian National Spatial Data Infrastructure[7]. These ontologies are represented in OWL and the Jena framework is used to parse them. After that, we gathered several spatial services (WMS and WFS) offered by several providers throughout Brazil. Each service was processed and their information was stored in a database. Besides, we registered information concerning each feature type they offer. Each type was semantically annotated through the domain ontologies defined in the infrastructure. Currently, this database stores about 457 feature types, distributed among 21 geospatial web services, from 16 service providers. All this information is stored in a PostgreSQL/PostGIS database server.

To illustrate the results obtained during the evaluation process, let us suppose a simple query in which the user wants to find feature types regarding *WaterCourses* in Brazil. In the database used for evaluation, there are 70 feature types directly related either to this concept or to one of its subclasses. These types are distributed among the services offered by 8 different Brazilian sources: the National Water Agency (ANA), the Executive Agency of Water Management of the State of Paraíba (AESA), the National Agency for Electrical Energy (ANEEL), the Brazilian Institute for Environment (IBAMA), the Ministry of Fisheries and Aquaculture (MPA), the Protection System for Amazon (SIPAM), the Department of Water Resources of the State of Santa Catarina (SIRHESC) and the Federal University of Minas Gerais (UFMG), according to the Table II. For each entry in the table, we have the provider name, the thematic relevance of the concept search (WaterCourse) to the service, the number of feature types annotated with the search concept, the number of feature types annotated with *WaterCourse* subclasses and the number of feature types annotated with concepts which are related to the search concept through a composition relationship.The inverse frequence value to the search concept is 0.419.

After executing the query, 36 feature types obtained a relevance degree greater than or equal to 90%. In this category were all types exactly annotated with the search concept. The feature types offered by the ANA and the MPA are listed first, with relevance around 91%. After, the result shows the 29 types of data provided by the SIRHESC with relevance around 89% and the types offered by the AESA, with relevance around 88%.

The second category contains feature types that have a relevance value between 76% and 86%. In this case there were two types of results. The first one contains data types that are associated with exactly the search concept, but are offered by services with low thematic relevance. In this case, we have data types offered by ANEEL, IBAMA and SIPAM. These types have gained importance around 86%, 86% and 85%, respectively. The second one is composed of data whose services have high thematic relevance, but they are annotated with concepts that represent subclasses of the search concept. Likewise, the other two feature types offered by ANA had relevance around 76%.

The third category of results includes 18 feature types that have a relevance value between 60% and 80%. This category contains data types offered by services in which the *WaterCourse* theme is highly relevant, but have been annotated with subclasses of the search concept. The remaining data types offered by SIRHESC, AESA, UFMG, ANEEL and SIPAM are listed, in this order. Finally, the last category contains the data types of services offered and AESA and UFMG that were annotated with concepts that have a composition relationship with the search concept. The relevance values were 58% and 51% respectively.

7. CONCLUSION AND FURTHER WORK

SDIs play an increasingly important role in the dissemination of geographic information offered by several organizations. However, locating geographic data offered by these infrastructures in an efficient and precise manner is still a hard task. Though ontology-based solutions have improved the discovering

[7]CONCAR: `http://www.concar.ibge.gov.br/arquivo/PlanoDeAcaoINDE.pdf`, in Portuguese

Table II. Data providers example concerning the hydrography concept

Provider	Thematic Relevance	Search Concept	Subclasses	Composition
ANA	1	2	4	0
AESA	0.562	3	6	1
ANEEL	0.162	5	5	0
IBAMA	0.125	3	0	0
MPA	1	4	0	0
SIPAM	0.051	2	2	0
SIRHESC	0.704	29	2	0
UFMG	0.307	0	3	1

process, there is still the need to evaluate the relevance of the retrieved results to the user, such that the more relevant results can be exhibited first.

This article described a framework that combines the notion of semantic similarity between concepts defined in ontologies and ideas applied to the classic information retrieval to evaluate how relevant are the spatial data offered by the infrastructure to an end-user's query. Though the present results have shown that the approach is interesting, some future works are still necessary.

One of the works necessary in the future consists in extending the notion of semantic similarity to treat more complex concepts and relationships, as, for example, concepts defined through conjunction, disjunction and negation of other concepts. Another important future work is to evaluate the user preferences once the result is presented. Along with validating our approach further, this work will allow a better adjustment of the weights used to calculate semantic similarity. Still, there is the need to evaluate the similarity between concepts defined in different ontologies, so that an even better recall for queries can be obtained.

REFERENCES

Athanasis, N., Kalabokidi, K., Vaiti, M., and Soulakellis, N. Towards Semantics-based Approach in the Development of Geographic Portals. *Computers & Geosciences* 35 (2): 301–308, 2009.

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

Baeza-Yates, R. A. and Ribeiro-Neto, B. A. *Modern Information Retrieval*. ACM Press / Addison-Wesley, 1999.

Berners-Lee, T., Hendler, J., and Lassila, O. The Semantic Web. Scientific American, 2001.

Bowers, S. and Ludäscher, B. An Ontology-Driven Framework for Data Transformation in Scientific Workflows. In *Data Integration in the Life Sciences*. Leipzig, Germany, pp. 1–16, 2004.

Egenhofer, M. J. Towards the Geospatial Semantic Web. In *Proceedings of the ACM International Symposium on Advances in Geographic Information Systems*. McLean, USA, pp. 1–4, 2002.

Guarino, N. Formal Ontology, Conceptual Analysis and Knowledge Representation. *International Journal of Human Computer Studies* 43 (5-6): 625–640, 1995.

Janowicz, K. Sim-DL: Towards a Semantic Similarity Measurement Theory for the Description Logic ALCNR in Geographic Information Retrieval. In *On the Move to Meaningful Internet Systems 2006: OTM 2006 Workshop*. Montpellier, France, pp. 1681–1692, 2006.

Janowicz, K., Wilkes, M., and Lutz, M. Similarity-Based Information Retrieval and its Role within Spatial Data Infrastructures. In *Proceedings of the International Conference on Geographic Information Science*. Park City, USA, pp. 151–167, 2008.

Klien, E., Lutz, M., and Kuhn, W. Ontology-Based Discovery of Geographic Information Services - An Application in Disaster Management. *Computers, Environment and Urban Systems* 30 (1): 102–123, 2006.

Kuhn, W. Geospatial Semantics: Why, of What, and How? In S. Spaccapietra and E. Zimányi (Eds.), *Journal of Data Semantics*. Springer, Berlin, pp. 1–24, 2005.

Lemmens, R., de By, R. A., Gould, M., Wytzisk, A., Granell, C., and van Oosterom, P. Enhancing Geoservice Chaining through Deep Service Descriptions. *Transactions in GIS* 6 (1): 849–871, 2007.

Lutz, M. Ontology-Based Descriptions for Semantic Discovery and Composition of Geoprocessing Services. *Geoinformatica* 11 (1): 1–36, 2007.

Lutz, M. and Klien, E. Ontology-Based Retrieval of Geographic Information. *International Journal of Geographical Information Science* 20 (3): 233–260, 2006.

Lutz, M. and Kolas, D. Rule-Based Discovery in Spatial Data Infrastructure. *Transactions in GIS* 11 (3): 317–336, 2007.

Lutz, M., Sprado, J., Klien, E., Schubert, C., and Christ, I. Overcoming Semantic Heterogeneity in Spatial Data Infrastructures. *Computers & Geosciences* 35 (4): 739–752, 2008.

Martin, D. L., Paolucci, M., McIlraith, S. A., Burstein, M. H., McDermott, D. V., McGuinness, D. L., Parsia, B., Payne, T. R., Sabou, M., Solanki, M., Srinivasan, N., and Sycara, K. P. Bringing Semantics to Web Services: The OWL-S Approach. In *Semantic Web Services and Web Process Composition: First International Workshop*. San Diego, USA, pp. 26–42, 2004.

Rada, R., Mili, H., Bicknell, E., and Blettner, M. Development and Application of a Metric on Semantic Nets. *IEEE Transactions on Systems, Man and Cybernetics* 19 (1): 17–30, 1989.

Smits, P. and Friis-Christensen, A. Resource Discovery in a European Spatial Data Infrastructure. *IEEE Transactions on Knowledge and Data Engineering* 19 (1): 85–95, 2007.

Stock, K. M., Atkinson, R., Higgins, C., Smal, M., Woolf, A., Millard, K., and Arctur, D. A Semantic Registry Using a Feature Type Catalogue instead Ontologies to Support Spatial Data Infrastructures. *International Journal of Geographic Information Science* 24 (2): 231–252, 2010.

Wiegand, N. and Garcia, C. A Task-Based Ontology Approach to Automate Geospatial Data Retrieval. *Transactions in GIS* 11 (3): 355–376, 2007.

Williamson, I., Rajabifard, A., and Feeney, M.-E. F. *Developing Spatial Data Infrastructures: From Concept to Reality*. Taylor & Francis, 2003.

Zhang, C., Zhao, T., Li, W., and Osleeb, J. P. Towards Logic-Based Geospatial Feature Discovery and Integration Using Web Feature Service and Geospatial Semantic Web. *International Journal of Geographic Information Science* 24 (6): 903–923, 2010.