Geographic Information Retrieval by Topological, Geographical, and Conceptual Matching

Felix Mata

PIIG Lab – Centre for Computing Research
National Polytechnic Institute
Av. Juan de Dios Bátiz s/n, 07738, México, D.F., Mexico
migfel@sagitario.cic.ipn.mx

Abstract. Geographic Information Science community is recognized that modern Geographic Information Retrieval systems should support the processing of imprecise data distributed over heterogeneous repositories. This means the search for relevant geographic results for a geographic query (Q_G) even if the data sources do not contain a result that matches exactly the user's request and then approximated results would be useful. Therefore, GIR systems should be centred at the nature and essence of spatial data (their relations and properties) taken into consideration the user's profile. Usually, semantic features are implicitly presented in *different* data sources. In this work, we use three heterogeneous data sources: vector data, geographic ontology, and geographic dictionaries. These repositories usually store *topological relations*, *concepts*, and *descriptions* of geographical objects under certain scenarios. In contrast to previous work, where these layers have been treated in an isolated way, their integration expects to be a better solution to capture the semantics of spatial objects. Thus, the use of spatial semantics and the integration of different information layers improve GIR, because adequate retrieval parameters according to the nature of spatial data, which emulate the user's requirements, can be established. In particular, we use topological relations {*inside, in*}, semantic relations {*hyperonimy, meronimy*}, and descriptions {*constraints, representation*}. An information extraction mechanism is designed for each data source, while the integration process is performed using the algorithm of ontology exploration. The ranking process is based on similarity measures, using the previously developed confusion theory. Finally, we present a case study to show some results of integrated GIR (iGIR) and compare them with Google's ones in a tabular form.

1 Introduction

Geographic Information Retrieval (GIR) is becoming increasingly popular task of using Geographic Information Systems (GIS). Due to the nature of geographic data, these are usually distributed over numerous heterogeneous repositories that makes the task challenging. Several proposals have been cited as methods to perform this task [25], but existing methodologies do not handle the variety of data sources in order to solve the problem adequately.

Usually, an approach is centered at just one of them; see e.g. [26]. In contrast, we believe that only an approach, integrating different information sources can essentially

F. Fonseca, M.A. Rodríguez, and S. Levashkin (Eds.): GeoS 2007, LNCS 4853, pp. 98–113, 2007.

improve GIR. The present paper is based on this belief, presenting a systemic approach to GIR.

Our approach consists of retrieving geographic information by processing queries, which can be split into a triplet <what, relation, where>, where *"what"* denotes a geographic object, *"where"* can be a spatial reference or a geographic object, and *"relation"* denotes a spatial relation linking *"where"* and *"what"*. These queries have been used in other works such as [23]. The approach is based on a retrieval strategy that uses three types of matching: the first one is a topological matching, i.e. topological relations extracted from overlaying data layers such as {*in, contain*}; the second one is a geographical matching, i.e. constraints obtained from dictionaries such as {*Airports represented by points or polygons*}; and the last one is a conceptual matching, given by a geographic ontology such as {*type of Airport*}. Thus, we use three heterogeneous data sources: vector files, dictionaries, and geographic ontology. A motivation to use these data sources is that they store different relations and properties depicting the nature of spatial data. These data sources have been also used in previous works [2][3][4] but separately. In contrast, we design herein an integrated system, which use all three sources, seeking for more powerful GIR. Hence, iGIR integrates a few processes (to be described in the following): querying, retrieving, and the integration and ranking. Figure 1 shows the framework of approach and the overall retrieval strategy.

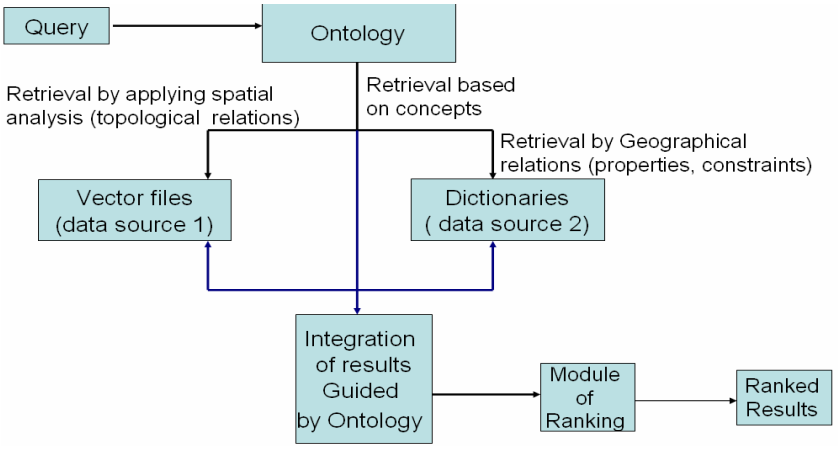

Fig. 1. Framework to retrieve geographic information

In the query processing, each data source allows associating geographic objects or spatial relations to each item of the query. This process starts by submitting a query into the system; the query is processed and all elements of the previously described triplet are identified. Then, a priori constrained ontology is explored to find the concepts, which correspond to the triplet's elements. The goal is to determine what other relations and objects should be required to be searched for and which data sources should be used; this is achieved by extracting context, where a context consists of

neighbour relations of a particular concept. Next, retrieval process is performed in the corresponding data source, and the answers are integrated into a set of results.

The final processes are the integration and ranking: a set of results is ranked according to their similarity using the confusion theory [1]. In essence, iGIR is based on integration of retrieval results: if a data source does not contain a relevant result, the other sources either provide an answer or, in the best case, each one adds a relevant answer to final set of results.

The first one of the data sources used in present research is an a priori and manually developed *ontology*. It contains the knowledge about geographic objects of a particular domain (e.g. *hotels in tourism domain*, *rivers in hydrology domain*). More generally, ontologies contain concepts and semantic relations between them {e.g. *meronimy*, *hyperonimy*}. In addition, a context is defined as the set of relations, which link a concept with other concepts. For example, in our ontology the concept *"Agave"* is linked to the concepts "*Jalisco"* and "*Plant"* by relations "*is-a"*, and *"grows"* respectively, thus the context = {*"is-a", "grows", "Plant", "Soil", "Weather", "Country"*}. Ontology is implemented in XML. The XML structure allows using or integrating other ontologies and, thus, our ontology can be enhanced. The systemic use of ontology in the retrieval process is described in section 3.1.

The second one of data sources is *vector files*. They are used to obtain the topological relations between data layers by means of overlaying operation (e.g. a layer of roads overlaid with airports, generated the new layer "*roads connect airports*"). In addition, other spatial relations can be discovered. To achieve this, a project (set of layers) is used: according to the parameters of the request, the appropriate layers or attributes are retrieved. A similar approach using spatial Bayesian learning is described in [3]. The processing of vector data is pointed out in section 3.3.

The last one of data sources is dictionaries. We use the dictionaries of INEGI-Mexico (National Institute of Statistics, Geography and Informatics). They contain descriptions, representations, scale, and constraints of particular objects (section 3.2).

Summing up, iGIR uses three data sources, described in above paragraphs, to retrieve geographic information by means of three matching.

The rest of the paper is structured as follows: Section 2 outlines related work. Section 3 describes the retrieval strategy: Sections 3.1 to 3.3 explain the mechanisms of conceptual, geographical, and topological matching, respectively. In addition, these sections describe the ontology design, the characteristics of dictionaries and vector data. In Section 4 some retrieval results are presented. Finally, in section 5 the conclusions as well as a future work are sketched out.

2 Related Work

To date, GIR presents several challenges; some of them have been treated using different approaches. For example, Rule-based methods and Data-driven methods are described in [5]; this article presents several heuristics to access data resources. Other proposed approach is a geographic search using a query-expansion [6]; the authors used a Google API. However, one of the serious disadvantages of this approach is that the query expansion (number of query terms) is constrained by the search engine.

Even if, it is possible to use an added term in order to disambiguate the words of query, this can also add more ambiguity. Thus, the retrieval process is not produced good results, if many terms are needed, because a number of terms required for disambiguation are a priori unknown.

Other proposals are focused on solving the problem of words ambiguity (words which describe geographic objects). The proposed solutions are based on a knowledge representation such as hierarchies of terms, taxonomies, and ontologies. Most of them use textual or syntactic properties; while others describe query processing, missing, however, spatial relations; see e.g. [7]. Inside this group of works, several semantic approaches have been also proposed; one of the main contributions consists of including ontologies and semantic annotation into the retrieval process; an example of such approaches is described in [8].

Although, the GIS community suggested and made emphasis on the use and treatment of spatial relations, only a few studies have been addressed these issues; see [9] [10]. A recent work focused on qualitative spatial reasoning; an example can be found in the often-cited model of topological relations between point sets [11]. Taking into account the above analysis of the state-of-the art, we use vector files, because they are very rich in spatial relations.

Ontologies [12] [13] have been widely used in several semantic approaches. They are now applied in many domains and in particular in GIS [14] [15]. Nevertheless, the proposed approaches do not consider processes and algorithms to explore ontologies. This would be, however, useful, because ontology describes domain theories and the explicit representations of the data semantics [16]. Thus, ontology can be used to discover the semantics of geographic objects involved in a query. Moreover, the algorithms to explore these ontologies and their semantics are required. Thus, we use herein ontologies and propose an algorithm to extract the semantics and domain knowledge stored into them.

On the other hand, many approaches in Information Retrieval (IR) are used the term-based Vector Space Model (VSM) [17]. They are based on lexicographic term matching. While, in iGIR the matching is performed by conceptual matching, topological relations, and descriptions of geographic objects according to the semantics of spatial data.

IR systems use models, techniques, and mechanisms to extract information that has already been processed and stored (e.g. plain text files, databases, XML files). In these systems, the fast processing of queries is possible, because the index structure has been previously built. The same idea is applied in GIR; see e.g. [18].

Besides this, the index structure is also used in domain dictionaries. Thus, we use dictionaries to extract properties and constraints of geographic object. These dictionaries are trusted and consensual sources, because they are designed by specialized and large institutions such as INEGI-Mexico or NASA-USA.

Our method is based on information retrieval guided by ontology, using *geographic queries*. For example, ontology describes where a plant grows, its type, and so on. Thus, we search for this plant either in dictionaries or in vector data. Next, the integration is guided by the relations between geographic objects. Finally, the ranking process is based on the confusion theory, measuring and controlling the dissimilarity between retrieved results. The overall system is described in the subsequent sections.

3 Strategy of Retrieval

For each data source, the goal consists of retrieving results according to semantic relations defined by ontologies, dictionaries, and vector files. Once the results are found, the integration is performed using spatial relations, and the final set of results is generated. Next, the ranking process is applied and the retrieval ends. We describe the overall retrieval process and the integration of three matching layers (conceptual, geographical, and topological) in the following subsections.

3.1 Conceptual Matching

This is the first step of the retrieval strategy, in which we use ontology. It plays the role of an expert in a specific domain, simulating the user's knowledge about this domain. Ontology allows guiding the retrieval, indicates which data should be searched for and where. In other words, ontology describes the way to retrieve relevant results according to semantic relations between geographic objects. For example, considering the following query Q_{G1} = {*"Hotels near Airport Benito Juarez"*} submitted by two types of users: a GIS user and a GIS neophyte. In both cases, the expected results are different: the GIS user wants to find digital data (vector files), while the neophyte wants to find the locations where the hotels stand near the Airport, and other information such as lodging prices, services, and so on. Varying the number of data sources used in our system, we can satisfy these two requirements. Moreover, we require knowing what type of data should be searched for. Thus, ontology defines the properties and relations of each geographic object (Geo_{obj}) i.e. it describes what is a Hotel, its type of representation, its properties and relations to other objects. Other geographic objects and relations involved in query are processed in the same way.

Ontology has manually built using articles from *Wikipedia*. The categories and links contained in each article have been considered as parameters to define relations and concepts of ontology. The semantic relations are classified according to their *meronimy* and *hyperonimy*. Wikipedia is a free online encyclopedia http://wikipedia.org/. In other works it has been used: 1) as data resource [27]; 2) for ontologies design [19]; 3) for words disambiguation [20].

Figure 2 shows a fragment (in Spanish) of the Wikipedia article and a fragment of the generated ontology, according to query Q_{G2} = {"*Agave grows Country*"}.

According to semantic relations between geographic objects, figure 2 depicts a fragment of ontology generated from the Wikipedia article (*Agave*). Note that *Jalisco* and *Plant* are linked through the Agave concept. Corresponding properties are extracted from words in bold and the relations are obtained from verbs, which link the concepts. In addition, the classes are defined and one of the properties is a list of synonyms. These synonyms are also extracted from the Wikipedia articles. In general, ontology is used to explore a data structure of the ontology tree. The goal is to find a matching for a particular concept. For example, assume that a user wants to know where the Agave grows, and submits the following query Q_{G2}= {*"Agave grows Country"*}. In the process of term's identification (triplet) the result is: ***what* = {Agave}; rel = {grows}; *where*= {Country}**. To find a matching for the overall query, we will classify it into four types: *atomic*, *partial*, *complete*, and *null*. A

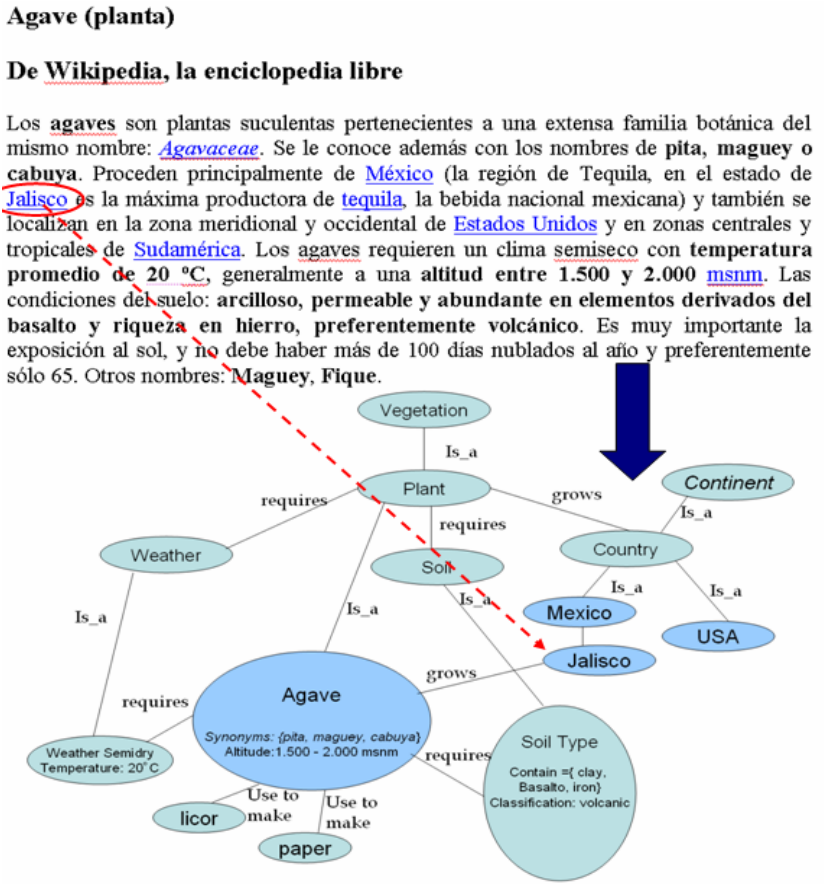

Fig. 2. Fragment of ontology generated from the Wikipedia article

matching is *atomic*, if one of the elements of query's triplet is identified in ontology. For example, if the object "*Agave*" is found, but neither class *"Country"* nor relation *"grows"* are found then we have an *atomic* matching. A matching is *partial*, if the geographical objects are not found, but relation or relations are presented in ontology. For example, if the class or object *"Country"* is not found, but the relation *"grows"* is found, then we have a partial matching. A matching is *complete*, if all three elements of query's triplet are identified in ontology. Finally, a matching is *null*, if none of the elements of query's triplet is identified. In this case, the retrieval system returns a Geo_{obj} that is processed by using the algorithm of confusion [1]. This algorithm measures the dissimilarity (confusion) for each element of triplet and returns the concept or relation somewhat close to the expressed term. For example, if a user searches for *"rivers"*, then the algorithm of confusion will return *"bodies of water"* as a farther matching, and *"lakes"* as a closer matching. Note that this process can be automatically controlled up to given error (confusion).

Table 1. Results of ontology exploration

Query (Q_{G1}): "Hotels near Airport Benito Juarez"	
Concept	***Geo_{obj} returned if ontology matching is complete***
Near	periphery , time, distance
Hotel	Tourism, lodging
Airport	hangar, transport
Query (Q_{G2}): "Agave grows Country"	
Concept	***Geo_{obj} returned if ontology matching is complete***
Agave	Plant, Desert Plant, Vascular Plant
Grow	Increase, change, develop
Country	Administrative district, USA, Mexico

Table 1 shows results if the matching is complete for the queries $Q_{G1=}$ {*"Hotels near Airport Benito Juarez"*} and $Q_{G2=}$ {*"Agave grows Country"*}.

1. **Begin**
2. Select each element of geographic query (Q_N) {what, rel, where}
3. Set the topological relation according to rel (Tr) (using association rules)
4. **Retrieving guided by ontology:**
5. Search (by concept name) the corresponding concept to geographic objects "what" and "where" into ontology
6. If (there_are_matches)
7. Then
   a. Extract the context {parent nodes, child nodes, neighbourhood nodes, and the instances for the main concept} by using semantic relations
   b. Extract the properties of concept ($P_{concept}$)
   c. Set context, relations and properties into array $Ont_{results}$
   d. Generate new queries (New_{Query})according to elements in $Ont_{results}$
8. Else
9. Request Q_N to Dictionary and vector Data
10. **Retrieving Dictionary Data:**
11. Request data to dictionary using values of neighbourhood nodes
12. Return response ($Resp_{Ont}$)
13. **Retrieving Vector Data:**
14. Using the response, request de vector data (layers)
    i. If (there_are_ matches (vector data))
    ii. Then
    iii. Request the corresponding relation (Tr)
    iv. Return $Resp_{Topological}$
15. **Integration and Ranking:**
    i. Set the $Resp_{Topological}$ and $Resp_{Ont}$ into final set (FS_{geobj})
    ii. Rank the FS_{geobj} and show to the user
16. Else
17. Select each element of geographic query (Q_N) {what, rel, where}
18. Search into dictionary the objects what and where
19. If (there_are_matches)
20. Then
21. Extract properties and relations
22. Set into array PR_{dict}
23. Return PR_{dict}
24. Else
25. Search into vector data, geographic objects *what* and *where*
26. Return the matches

Fig. 3. Algorithm of ontology exploration used in the retrieval strategy

Table 1 also shows several relations and geographic objects for concepts found by the algorithm of ontology exploration. We use this algorithm to search for the relevant ontology concepts and relations, and then apply the matching according to the submitted geographic query. A fragment of algorithm is described in Figure 3.

Figure 3 describes the steps to process the query, where we explain the functionality and processing using the query Q_{G2} = {"*Agave grows Country*"}. According to the algorithm, each triplet's element is identified, and for the relation *grows*, the overlay spatial operation is applied. This operation is defined according to a set of rules. These are described in section 3.2. Then results for the lines *5* to *7d* are:

Context= {
Parents (Plant);
Neighbourhood (soil, weather, country)
}

Using the context, new queries are generated by combining the elements of context. Thus, the generated queries are the following:

Q_{G3}= *{Agave grows Mexico}*
Q_{G4}= *{Agave grows USA}*
Q_{G5}= *{Plants grows Mexico}*
Q_{G6}= *{Plants grows USA}*

The queries Q_{G3} to Q_{G6} are searched for into the dictionary and vector data. If a term of query is found, then its properties, constraints, representation, and relations are integrated into a set of results. In the worst case, there is no matching for an element of query. In this case, the queries are submitted and processed by the confusion module, where the new queries are generated and resubmitted to the initial process. Therefore, in a successful scenario, each object is found according to the previously established criteria (see lines 4-9 of Figure 3), in which each one of them are requested in the source of vector data which fulfil these criteria. Sections 3.2 and 3.3 are described in detail the rest of the process (geographical and topological matching).

3.2 Geographical Matching

The next step consists of making a geographical matching. This process uses as a data source the information dictionaries. They represent a consensual agreement between the GIS specialists and contain the scale, properties, constraints, and relations, etc. of geographic objects. The dictionaries are initially in *PDF* afterwards to be semiautomatically transformed into XML files, using the API, *PJX* (see http://java-source.net/open-source/pdf-libraries/pjx). Figure 4 shows a fragment of dictionary in PDF format and the corresponding XML file. In particular, the fragment describes the object *Airport*. Due to didactical reasons and available data, we explain the process of geographical matching using a query which contains airports.

Figure 4 depicts the sections of dictionary; each section is to be transformed into nodes of XML file. Now, we explain the process to extract the information required to improve GIR. Consider the generated queries of section 3.1 (Q_{G3} to Q_{G6}) or original

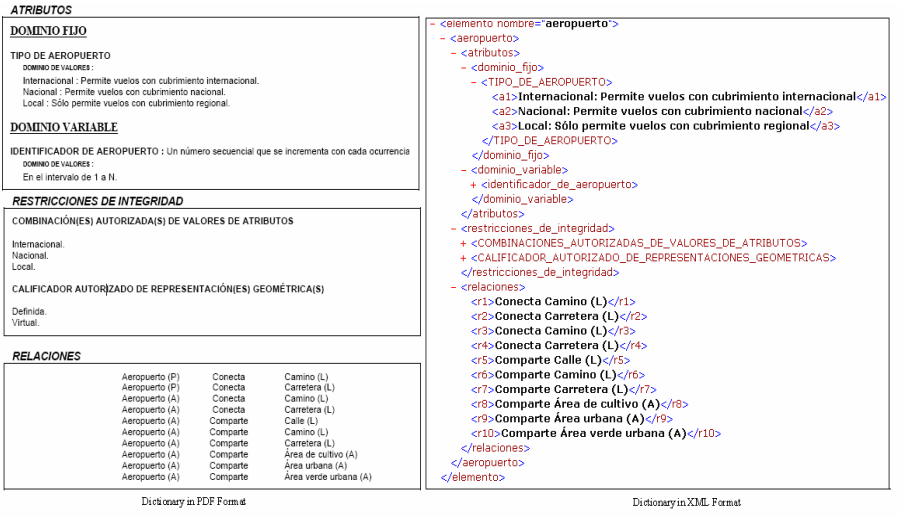

Fig. 4. INEGI dictionaries used in the geographical matching

query Q_{G2}. In this case, due to availability of data, we use the query Q_{G1}= {*"Hotels near Airport Benito Juarez"*}. This query is processed to obtain the triplet: what = {*"Hotels"*}; relation = {*"near"*}; where = {*"Airport Benito Juarez"*}. These properties and relations are extracted from the dictionary for the object "*Airport*", where the results are: Relations = {"*connect*", "*sharing*"} Properties= {Type = {"*Local*", "*National*", "*International*"}} Constraints= {primitive of representation = {"*point*", "*polygon*"}}. The object *Hotel* has no occurrences in the dictionary, therefore, the process continues by using only the object *Airport*. The constraints are needed because of possible semantic changes. For instance, if an airport is depicted by a point feature, then it represents a building of operations, while if it is depicted by a polygon, then it represents the area, infrastructure and services of air navigation. Additionally, we define a set of rules for processing each relation. These rules are established in our previous work [24]. The main idea is to associate a topological relation to the relation expressed in query (e.g. "*near*" is associated to relation "*connect*"). In this case one of the rules of association to relation "*near*" is the following:

$\mathbf{R_{1\ (NEAR)}}$ = **{X near Y, if X and Y are connected by Z}** where X, Y, are geographic objects (Geo_{Obj}), represented by points or polygons, while Z is a geographic object represented by an arc. Moreover, its length is less than 1 kilometer (e.g. Z is a road). Finally, *"connected"* is a relation between X and Y.

Therefore, the retrieval in this step consists of searching for documents which fulfill the above rule. Therefore, the parameters of searching (P_{search}) are:

$\mathbf{P_{search1} = \{Geo_{ObjX}\ connects\ Geo_{ObjY}\}}$ where Geo_{ObjY} can be a point or a polygon.
$\mathbf{P_{search2} = \{Geo_{ObjZ}\ sharing\ Geo_{ObjW}\}}$ where Geo_{ObjY} can be a point or a polygon.

These parameters represent the search performed by using vector files. This is the point where the next step (topological matching) starts. This process is described in section 3.3.

3.3 Topological Matching

The process of topological matching is based on topological relations between geographic objects. This process uses vector files as a data source. The data of this representation model is provided by INEGI and SCT (Secretary of Communications and Transportation). The data are processed to obtain a proprietary format file called herein *Topologyfile*. In particular, these files store the topological relations between two geographic objects. Table 2 shows the structure of a *Topologyfile* (File.geo).

Table 2. Structure of *Topologyfile*

ID	Id_GeoObj_1	LAYER_BELONG	Id_GeoObj_2	LAYER_BELONG	RELATION
1	5	Airport.geo	2	Roads.geo	C
2	10	Hotels.geo	2	UrbanArea.geo	S

C= Connects, S= sharing

Table 2 shows an identifier for each record which fulfils $\mathbf{P}_{\mathbf{search1}}$ and $\mathbf{P}_{\mathbf{search2}}$ (section 3.2). The following columns allow identifying each geographical object and its corresponding layer. The attributes indicate the type of relations between two geographical objects (e.g. "*connect*", *"inside"*). These relations are obtained by applying the overlay spatial operation (e.g. overlaying *roads* and *airports* = road A connects airport B). The details on how to get spatial relations into tables are described in [21]. Therefore, for the query Q_{G1}, the sources which include layers: *"Airports"*, *"Hotels"* and *"roads"* are explored to find that "*Airport Benito Juarez*" is connected by several streets and avenues, and some Hotels are also connected by the same streets and avenues. Then, the objects linked by the topological relations ("*connect*", *"sharing"*) are retrieved. Figure 5 shows a table in which an example of topological matching is presented.

ID	ID_OBJ_1	LAYER_BELONG	ID OBJ_2	LAYER_BELONG	RELATION
0	5	Airport.geo	2	Roads.geo	C

ID	ID_OBJ_1	LAYER_BELONG	ID OBJ_2	LAYER_BELONG	RELATION
12	7	Hotels.geo	2	Roads.geo	C

Fig. 5. Topological matching: Object 5 and object 2 are connected. Object 2 is connected to object 7.

Figure 5 shows how a result, where object 5 (*Airport Benito Juarez*) represented by a point, containing a relation *"connect"* to object 2 (*Avenue* "*Circuito Interior*"), is retrieved. Then, the next step is to find Hotels, where object 2 appears (in the best case, it will connect by the same relation). This way, we find that object 7 (*Hotel Holiday Inn*) has a relation to object 2 (*Avenue* "*Circuito Interior*"). This means that

Hotel Holiday Inn is connected by Avenue Circuito Interior and the Avenue is connected to the Airport Benito Juarez. These results are subsequently submitted to the integration and ranking module.

Finally, each result is integrated to a final set of results and submitted to the ranking module. Here the retrieval process ends and the ranking process starts. The ranking process is based on the similarity measures between concepts and relations called in [1] *confusion*.

4 Experiments

In this section, we present some screenshots of the query Q_{G1} = {"*Hotels near Airport Benito Juarez"*} processing, applying the algorithm of ontology exploration that returns the list of classes and objects to be searched for into dictionaries and vector files. In addition, the values of properties for each geographic object are visualized. These results are addressed to GIS neophytes. Figure 6 shows the result for the class Airport.

Figure 6 shows a retrieved document with the properties and attributes, which define the Airport class. The definition is based on Wikipedia documents (in Spanish)

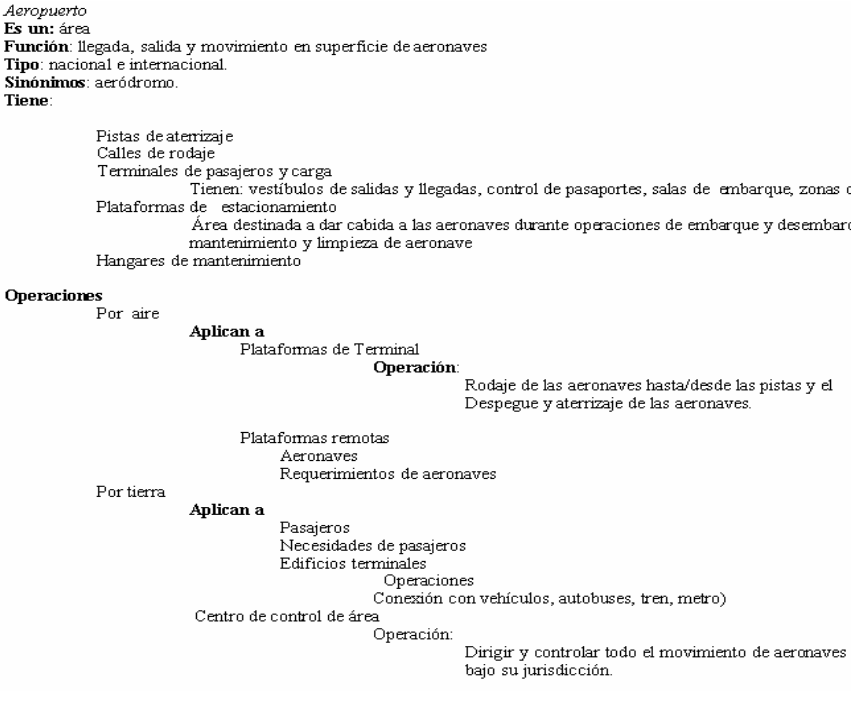

Fig. 6. The result of the algorithm of ontology exploration searching for the Airport class; this result is addressed to GIS neophytes

which contain the NL commonalities and popular use of geographic objects such as Airport. Therefore, these documents can be useful for travelers, businessmen, etc.

According to the algorithm of ontology exploration the contextual results are sent to the module of geographic matching. The goal is to find other objects related to the original query according to spatial relations. These results are addressed to the GIS specialists. Figure 7 shows the results of conceptual and geographical matching.

Figure 7 also shows the relations retrieved for the object "*Airport"* from dictionaries. In this case, some relations are: {*connect* and *sharing*} and some associated classes are: {"*highway"*, "*street"*, "*urban area*", etc}. The classes are represented as a

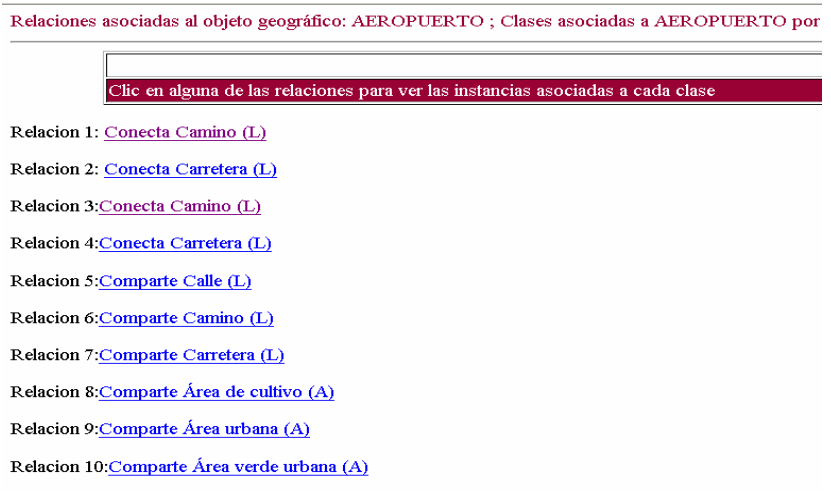

Fig. 7. Results of the query Q_{G1} = {"*Hotels near Airport Benito Juarez*"} processing by using dictionaries and guided by ontology

Aeropuerto

OBJECTID	TIPO	ENTIDAD	FC	CAPA DATOS (LAYER)
1	Internacional	AEROPUERTO	643	airport1.shp
2	Internacional	AEROPUERTO	643	airport2.geo
3	Internacional	AEROPUERTO	643	airport1.shp
4	Internacional	AEROPUERTO	643	airport1.shp
5	Internacional	AEROPUERTO	643	airport2.geo
6	Internacional	AEROPUERTO	643	airport2.geo
7	Internacional	AEROPUERTO	643	airport2.geo
8	Internacional	AEROPUERTO	643	airport2.geo
9	Internacional	AEROPUERTO	643	airport3.geo
10	Internacional	AEROPUERTO	643	airport3.geo
11	Internacional	AEROPUERTO	643	airport3.geo
12	Internacional	AEROPUERTO	643	airport3.geo
13	Internacional	AEROPUERTO	643	airport3.geo

Fig. 8. Results of the query Q_{G1}= {*"Hotels near Airport Benito Juarez"*} processing by using dictionaries, vector files, and guided by ontology

link, because they are related to instances of these classes. These instances are retrieved from vector data (e.g. searching for *highways, streets* with a relation such as *connect*, *sharing*).

Figure 8 shows the final results by selecting one of the links (shown in figure 7) for Airports. In this case, the relation is "*connect*".

The last column depicts the vector file associated to each object. In addition, this process generates a KML file if available vector data contain the latitude and longitude of spatial objects. Thus, figure 9 shows the KML file generated for the query Q_{G1}= {"*Hotels near Airport Benito Juarez*"}. Figure 9 also shows the area where the airport Benito Juarez is located. In addition, the subway stations near to it are displayed. That is why the area extent is larger than the original one.

Fig. 9. The KML file generated for the query Q_{GI} = {"*Hotels near Airport Benito Juarez*"}, using dictionaries, vector files, and guided by ontology

Finally, we test our approach comparing it with the results provided by *Google*. The results have been validated by the group of postgraduate students of the PIIG Lab. In general, the results have matched their expectations. In our test, the results of retrieval process have been classified into three types: *somewhat relevant, relevant, and irrelevant*. Relevant and irrelevant results are eloquent, while a result is somewhat relevant if it either contains a property of the geographic objects or an object related to them. For

Table 3. Statistics of final results

Geographic query	System Used	Somewhat relevant	Relevant	Irrelevant
1	*Google*	2	5	1
	iGIR	3	7	3
2	*Google*	0	2	5
	iGIR	6	7	2
3	*Google*	2	3	4
	iGIR	6	5	2
4	*Google*	1	2	4
	iGIR	4	5	2

example, retrieving "*Agave*" is somewhat relevant result, if "*tequila*" concept is searched for. This classification is based on the confusion measures [1].

Table 3 shows the experimental results, in which the number in columns *somewhat relevant*, *relevant*, and *irrelevant*, represents the number of documents, which have been found. The results are satisfactory and generally match the user's expectations. Nevertheless, additional tests are required, using other data and methods to measure the relevance of results. These are another part of work in progress. Finally, the ranking process is applied, using a set of previously defined geographical objects.

5 Conclusions

This work describes an approach to perform geographic information retrieval based on integration of three sources of geographic information (iGIR system). The main idea is to extract and process the properties and relations of the geographic objects which appear in the data sources (the former store descriptions, constraints, topological and geographic relations). The approach is based on the algorithm of ontology exploration. A method to match the concepts of geographic objects by their relations and properties, not only syntactically but also semantically, is developed as well. The retrieval is guided by ontology. It is manually designed and based on Wikipedia articles. Wikipedia is a free online encyclopedia – a trusted and consensual information resource also used in other GIR works. Ontology helps to decide where and what should be searched for into other two data sources – geographic dictionaries and vector files, thus, simulating the user's judgement. INEGI-Mexico dictionaries have been used in this work. These contain descriptions, properties, and relations of particular geographic objects at certain scale. Vector files are used in form of a proprietary file format called herein *Topologyfile*. These files represent topological relations such as *adjacent*, *in*, etc. This work is primarily different from others (e.g. query expansion), because the geographic information retrieval is made by matching concepts using the algorithm of ontology exploration, and their integration with geographic dictionaries and vector data. The purpose of integration is to provide adequate search parameters and in consequence improve the overall retrieval process.

The paper exposes some results of processing the geographical queries over heterogeneous repositories. The retrieved results are addressed to two types of users: GIS specialists and GIS neophytes according to either their requirements or their profiles. Nevertheless, additional testing is needed to validate the overall approach as well as its components; especially on large document collections (e.g. the test of the ranking module requires such large collections, while present test used only 20 documents).

Acknowledgments

The author of this paper wishes to thank his scientific advisor Dr. Serguei Levachkine, the Centre for Computing Research (CIC), SIP-IPN, National Polytechnic Institute (IPN), and the Mexican National Council for Science and Technology (CONACYT) for their support.

References

1. Levachkine, S., Guzman-Arenas, A.: Hierarchy as a New Data Type for Qualitative Variables. Expert Systems with Applications: An International Journal 32(3), 899–910 (2007)
2. Gazetteer Development at the Alexandria Digital Library Project. available at: http://www.alexandria.ucsb.edu/gazetteer
3. Walker, A.R., Pham, B., Moody, M.: Spatial Bayesian Learning Algorithms for Geographic Information Retrieval. In: Proceedings of the 13th annual ACM international workshop on Geographic information systems, GIS 2005, 42, Bremen, Germany (2005)
4. Clough, P.: Extracting Metadata for Spatially-Aware Information Retrieval on the Internet. In: GIR 2005. Proceedings of the Workshop on Geographic Information Retrieval, Bremen Germany, pp. 25–30. ACM Press, New York (2005)
5. Egenhofer, M.: Interaction with Geographic Information Systems via Spatial Queries. Journal of Visual Languages and Computing 1(4), 389–413 (1990)
6. Delboni, T.M., Borges, K.A., Laender, A.H., Davis, C.A.: Semantic Expansion of Geographic Web Queries Based on Natural Language Positioning Expressions. Transactions in GIS 11(3), 377–397 (2007)
7. Maedche, A., Staab, S., Stojanovic, N., Studer, R., Sure, Y.: SEAL A Framework for Developing Semantic Web Portals. In: Read, B. (ed.) Advances in Databases. LNCS, vol. 2097, Springer, Berlin (2001)
8. Arpinar, I.B., Sheth, A., Ramakrishnan, C., Usery, L., Azami, M., Kwan, M.: Geospatial Ontology Development and Semantic Analytics. Transactions in GIS 10(4), 551–576 (2006)
9. Heinzle, F., Kcopczynsky, M., Sester, M.: Spatial Data Interpretations for the Intelligent Access to Spatial Information in the Internet. In: Proceedings of the 21st International Cartographic Conference, Durban, South Africa (2002)
10. Shilder, F., Versley, Y., Habel, C.: Extracting Spatial Information: rounding, Classifying and Linking Spatial Expressions. In: Proceedings of the ACM SIGIR workshop on Geographic Information Retrieval, Sheffield, UK (2004)
11. Egenhofer, M., Franzosa, R.: Point-Set Topological Spatial Relations. International Journal of Geographical Information Systems 5(2), 161–174 (1991)
12. Guarino, N.: Formal ontology and information systems. In: Proceeding of FOIS 1998, Trento, Italy, pp. 3–15. IOS press, Amsterdam (1998)
13. Gruber, R.: A translation approach to portable ontology specifications. Knowledge Acquisition 5(2), 199–220 (1993)
14. Harding, J.: Geo-ontology Concepts and Issues, Report of a workshop on Geo-ontology, Ilkley UK (2003)
15. Hammond, B., Sheth, A., Kochut, K.: Semantic Enhancement Engine: A Modular Document Enhancement Platform for Semantic Applications over Heterogeneous Content. In: Kashyap, V., Shklar, L. (eds.) Real World Semantic Web Applications, pp. 29–49. IOS Press, Amsterdam (2002)
16. Koo, S., Lim, S., Lee, S.: Building ontology based on hub words for information retrieval. In: Proceedings of the IEEE/WIC International Conference on Web Intelligence, IEEE Computer Society, Los Alamitos, 466 (2003)
17. Baeza-Yates, R., Ribeiro-Neto, B.: Modern Information Retrieval. ACM Press, Addison-Wesley, New York (1999)
18. Jones, C.-B., Abdelmoty, A.-I., Finch, D., Fu, G., Vaid, S.: The Spirit Spatial Search Engine: Architecture, Ontologies and Spatial Indexing. In: Proceedings of Third International Conference, Maryland, USA. LNCS, vol. 3234, pp. 125–139. Springer, Berlin (2004)

19. Buscaldi, D., Rosso, P., García, P.: Inferring geographical ontologies from multiple resources for geographical information retrieval. In: Proceedings of 3erd Int. SIGIR Workshop on Geographic Information Retrieval, SIGIR, Seattle, pp. 52–55. ACM Press, New York (2006)
20. Martins, B., Silva, M.J., Silveira, M.: Challenges and Resources for Evaluating Geographical IR. In: Proceedings of the 2005 workshop on Geographic Information Retrieval, pp. 65–69. ACM Press, New York (2005)
21. Martinez, M.: Topologic Descriptor for Topographic Maps, Master of Science Thesis, Mexico, PIIG Lab, Centre for Computing Research, in Spanish (2006)
22. Wordnet: A Lexical Database for the English Language. available at http://www.wordnet.com
23. Martins, B., Silva, M., Freitas, S., Afonso, A.: Handling, Locations in Search Engine Queries. In: Proceedings of GIR 2006, the 3erd Workshop on Geographic Information Retrieval (2006)
24. Mata, F., Levachkine, S.: Semantics of Proximity in Locative Expressions. In: Rodríguez, M.A., Cruz, I., Levashkin, S., Egenhofer, M.J. (eds.) GeoS 2005. LNCS, vol. 3799, Springer, Heidelberg (2005)
25. Yee, W.G., Beigbeder, M., Buntine, W.: SIGIR06 Workshop Report: Open Source Information Retrieval systems (OSIR06). SIGIR Forum 40(2), 61–65 (2006)
26. Petras, V., Fredric, C.G., Larson, R.: Domain-Specific CLIR of English, German and Russian Using Fusion and Subject Metadata for Query Expansion. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.) CLEF 2005. LNCS, vol. 4022, pp. 226–237. Springer, Heidelberg (2006)
27. Cardoso, N., Martins, B., Chaves M., Andrade, L., Silva, M.: The XLDB group at GeoCLEF 2005. In: GeoCLEF 2005 Workshop, Poster Session (2005)