Graph-based Navigation Strategies for Heterogeneous Spatial Data Sets

Andrea Rodríguez[1] and Francisco Godoy[2]

[1] Department of Computer Science, University of Concepción
Edmundo Larenas 215, 4070409 Concepción, Chile
andrea@udec.cl
[2] Center for Oceanographic Research in the Eastern South-Pacific
FONDAP-COPAS, University of Concepción
P.O. Box 160-C, Concepción, Chile
fgodoyf@udec.cl

Abstract. Querying heterogeneous spatial databases involves not only characterizing and comparing the information content of several databases, but also *navigating or accessing* the data sets with the query answer. This work proposes a formalism that relates the information content of data sets by three basic types of *correspondence relations*: *data equivalence*, *difference of data omission*, and *difference of data commission*. These correspondence relations define the *information space* over which a navigation process is carried out. Based on a complete or an incomplete information space, this work proposes strategies that optimize the retrieval process of information coming from different databases. The results of this study show the advantages of defining the information space to select and access databases. In particular, strategies that estimate the information contribution of data sets based on correspondence relations outperform a strategy that considers a random list or a list of data sets sorted by size.

1 Introduction

In presence of syntactic, schematic, and semantic differences [15], accessing distributed databases requires an important effort to compare user queries with data stored in diverse databases. Most research in this area has focused on establishing the relationship among databases; that is, similarity, difference, and equivalence between database schemas [8] [7] [12] [17] [9]. A subsequent, but not less important, issue in querying and accessing heterogeneous databases deals with the way users *navigate* among data sets that contain the query answer. Users are typically presented with only a ranked list of databases that totally or partially satisfy their requests and then, they have to decide which of these databases they will access.

This work considers the process of accessing different databases that contain desired data as a problem of information navigation. For example, consider a user who wants to retrieve data of utility networks in a given urban area. These data may be obtained from different databases, each of them containing partial data of utility networks. The query to each database may result in different data sets (i.e., sets of tuples in relational databases), and users have to decide what data sets and in which sequence should be accessed. We claim that is not only important to know what databases contain the desired data, but to provide with strategies to organize the results of the query so that an efficient retrieval process is accomplished. In this context, efficiency refers to getting all or most information (i.e., new data) with the access to few databases.

The contribution of this work relates to the navigation in an information space composed of heterogeneous spatial data sets. This can be done after database schemas or metadata have been compared and databases that contain the desired type of data have been detected. This work proposes a formalism that allows one to relate spatial databases in terms of the equivalences and differences between their data sets.

We define an information space by characterizing the correspondences between data sets with three basic categories: *data equivalence*, *difference of data omission*, and *difference of data commission*. Using these categories in a graph-base representation, this work describes strategies that optimize the navigation of data sets with the query results. The objective of this optimization is to retrieve data that contribute to the answer, minimizing duplications while increasing the data retrieved. The work concentrates on spatial objects represented by regions. At an abstract level, the same approach may be useful in other domains of information.

The structure of the paper is as follows. Section 2 gives a brief review of related work. Section 3 describes the modeling of the information space with the correspondence relations between data sets and the graph-based representation. Section 4 presents the different strategies for information navigation, and Section 5 shows the implementation of the strategies. Final conclusions and future research directions are given in Section 6.

2 Related Work

In the area of information systems, the concept of information navigation has been associated with the visualization of retrieval results. In a retrieval process with heterogeneous data collections that totally and partially satisfied a user query, the problem becomes to select and browse the query results. The general idea for solving this problem is to use overviews of the information that may guide users in the retrieval process [2]. We consider the process of selecting and browsing documents as a process of information navigation, where two important approaches are:

- Category hierarchies. This approach assigns metadata to data sets based on their categories, which are then organized into a hierarchy. The problem with this strategy is that one could often need to look at a large collection of tags and, when the category has been found, search within the category. Examples of this type of approach are the computer classification system of the ACM

[1], which has developed a hierarchy of 1200 categories or the the popular search site Yahoo [19], which organized documents in many categories.

- Clustering techniques. There are several cluster-based solutions for helping users in searching and navigating data sets. They attempt to display overview information derived from the metadata or the extraction of common features in a collection. Then, clusters group data sets based on the similarity to one another. An early contribution is the paradigm Scatter/Gatter [3], where users are provided with a summary of the documents that have been clustered. More recent works have also considered dynamic clusters [20] and clustering and summary of query results for information navigation [14]. An example of such approaches is the web site VIVISIMO [16], which dynamically organizes the query results by topics.

A second perspective of information navigation relates to the process of accessing heterogeneous data sources. In this context, the navigation is carried out over a semantic structure, called information space, that connects different sources [13]. This work concerns the second perspective of a navigation process. It defines a structure upon which different strategies for selecting the path in the navigation process can be taken.

3 The Information Space

The information space describes the content relations between data sets. These relations are basic correspondence categories upon which a graph representation is defined to be used in the navigation process.

3.1 Types of correspondences between data sets

Three types of basic correspondence relations between data sets are: *equivalence*, *difference of commission*, and *difference of omission*. The distinction between *difference of commission* and *difference of omission* resembles the notions of *error of commission* and *error of omission*, respectively, associated with types of inaccuracy in an ontology of imperfection for information integration [18]. In such ontology, inaccuracy is the lack of correlation with the actual state of affair. Our notions of *difference of commission* and *difference of omission*, however, do not relate to errors. Commission and omission are here converse concepts related to the presence or absence of data, respectively. Since this work assumes no further information about the origin of data, all data sets are considered to equally contribute to the integration or retrieval of information. Thus, solving conflicting geometric representations (e.g., two different geometries for a same object) is a process carried out after all data are retrieved.

A definition of the correspondence relations follows.

- *Equivalence*. Considering the same thematic layer, two objects are said to be equivalent if they occupied the same space or have the same representative set of coordinates. We relax the definition of equivalence for not totally equivalent objects by considering that two objects from different databases and a same thematic layer are represented by equivalent and different regions. The equivalent regions are the intersection of the geometric representations of objects, whereas the different regions are the non intersecting representations of objects. This difference between objects is further classified into *difference of commission* and *difference of omission*.

Let x, y, z be regions belonging to objects from different data sets, then the equivalence relation between x and y ($x \equiv y$) satisfies the following basic properties:

$$\forall \quad x, y \quad [x \equiv y \rightarrow y \equiv x] \; \textit{symmetry} \tag{1}$$

$$\forall \; x, y, z \; [(x \equiv y \wedge y \equiv z) \rightarrow x \equiv z] \; \textit{transitivity} \tag{2}$$

We define an *Equivalent Operator* between two data sets A and B ($E(A, B)$) to be the equivalent regions between the data sets. The total area of these equivalent regions will be expressed by $|E(A, B)|$.

- *Difference of commission.* The difference of omission between two data sets A and B ($C(A, B)$) corresponds to the regions that are in A and not in B.
- *Difference of omission.* As the converse relation of $C(A, B)$, the difference of omission between data sets A and B ($O(A, B)$) corresponds to the regions that are in B and not in A.

Figure 1 presents the basic cases of the three types of correspondence between data sets A and B when considering the same spatial window ω. In this case, features a and c are equivalent in both data sets, feature b is in data set A but not in B; that is, there is a difference of commission of feature b between data sets A and B, and a difference of omission of feature b between data sets B and A. Likewise, d is a feature in data set B and not in data set A.

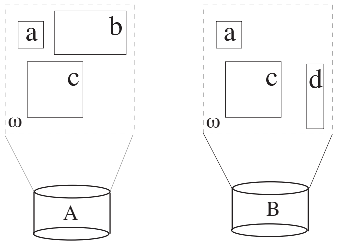


Fig. 1. Example of equivalences and differences between data sets

In the case of partial equivalence and matching of composite objects between two data sets A, B, the operators $|E(A, B)|$, $|C(A, B)|$ and $|O(A, B)|$ define the percentages of areas that are equivalent and different, without indicating how many objects are really equivalent or different. A further study could analyze when an object is equivalent or not, a problem that has been addressed in studies of data integration [18] [6] and consistency at multiple representations [5] [10] [4].

Let a and b be the total areas of objects in data sets A and B, respectively, the correspondence operators satisfy the following properties:

$$a = |E(A, A)| \tag{3}$$

$$a = |E(A, B)| + |C(A, B)| \tag{4}$$

$$a = |E(A, B)| + |O(B, A)| \tag{5}$$

$$b = (a - |O(B, A)|) + |O(A, B)| \tag{6}$$

3.2 Graph-based representation of the information space

The set of data sets with their correspondence relations can be seen as a directed and labeled graph, with nodes being data sets and edges being the correspondence categories between data sets. The graph can be simplified by using only one directed edge between nodes, since the inverse directed edge can be determined by the converse properties $|C(A, B)| = |O(B, A)|$.

For n data sets, a complete graph is of order $O(n^2)$. Under incomplete information, the composition of correspondence relations allows one to constrain the possible values of unknown information. Thus, a composition of correspondence relations between data sets A and B and between B and C indicates possible values for the correspondence relations between A and C.

Proposition 1. *Given three different data sets A, B, and C, with total areas of objects a, b, and c, respectively, and where known information are the results of the equivalent operator between A and B ($E(A, B)$) and between B and C ($E(B, C)$), the value of $|E(A, C)|$ has lower and upper bounds given by*

$$\forall A, B, C[max(0, |E(A, B)| + |E(B, C)| - b) \leq |E(A, C)| \leq min(a, c)] \tag{7}$$

Proof. A lower bound of equivalent areas between data sets A and C is determined by the minimum area of regions that are in common between regions in $E(A, B)$ and $E(B, C)$. These regions must be also in the equivalent regions between A and C by the transitivity property of equivalence (Equation 2). The minimum area can be obtained from the maximum area of regions that are different between the equivalent regions $E(A, B)$ and $E(B, C)$. In order to have a region in $E(A, B)$ that is not in $E(B, C)$, the equivalent region in $E(A, B)$ must be in $C(B, C)$. Thus, we cannot have more regions that are different between $E(A, B)$ and $C(B, C)$. Once $|C(B, C)|$ is less than $|E(A, B)|$, the lower bound of $|E(A, C)|$ is equal to $|E(A, B)| - |C(B, C)|$. Once $|C(B, C)|$ is larger than $|E(A, B)|$, the lower bound of $|E(A, C)|$ is equal to zero, since it assumes that all equivalent regions between A and B are also in the set of regions that are in B and not in C.

An upper bound of the area of equivalent regions between data sets A and C is given by the minimum area of the space used by objects in each data set, since one cannot have an area of equivalent regions larger than the total area of objects in the data sets.

Corollary 1. *Given that* $|C(A,C)| = a - |E(A,C)|$ *and* $|O(A,C)| = c - |E(A,C)|$, *then*

$$\forall A, B, C[a - min(a,c) \leq |C(A,C)| \leq a - max(0, |E(A,B)| - |C(B,C)|)] \quad (8)$$
$$\forall A, B, C[c - min(a,c) \leq |O(A,C)| \leq c - max(0, |E(A,B)| - |C(B,C)|)] \quad (9)$$

4 Information Navigation

The navigation strategies proposed in this paper aim at retrieving all data with a minimum effort or cost. Effort or cost in this case are considered to be the total amount of retrieved data, measured as the total area of objects retrieved. In such retrieval process, duplication may occur and the idea is to obtain all data, while accessing the minimum number of databases or accessing databases in a sequence that obtains most data as quick as possible. For all data we mean all different regions, since these regions are the real information contribution of data sets. If one of the data sets retrieved from a database contains all regions (objects) in a desired geographic window, the system should just access this database to obtain a complete answer. In an imperfect world, however, databases are incomplete. In such cases, to obtain a complete answer means to access more than one database, but hopefully, not all of them.

4.1 Navigation Strategies

There are different strategies retrieving all data from an information space. To illustrate the different alternatives, consider a simple case of four data sets (coming from different databases) (Figure 2), whose correspondence relations are codified by the triple $(|O()|, E()|, |C()|)$ as labels of the directed graph, and where the total area of objects in each data set is codified by the value in each corresponding node.

Common to all strategies is that they favor data sets with large objects' geometries. It is not the number of objects, but the objects that occupy more space what will guide the navigation strategy.

- *Size-based strategy.* A size-based strategy sorts the data sets based on the total area of objects in the sets, retrieving, for example, the data set in decreasing order of area. With this strategy, all data sets are retrieved. In the case of the example in Figure 2, the navigation sequence is given in Table 1.
- *Forward-based strategy.* A forward-based strategy takes the current position of the navigation (initially, the node of the data set with the largest area)

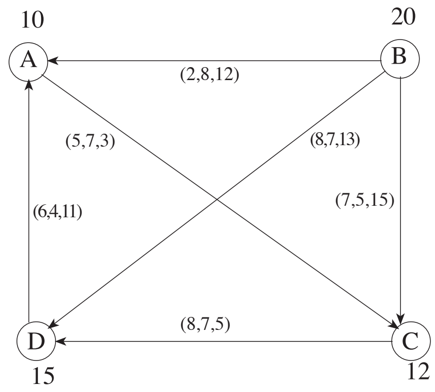


Fig. 2. Complete graph with four data sets

Data Set	Total area	Sequence order
A	10	4
B	20	1
C	12	3
D	15	2

Table 1. An example of navigation with the size-based strategy

and selects a next data set that has the largest area of commission with respect to the current data set, or inversely, a data set respect to which the current data set has the largest area of omission. Consider the example in Figure 2 and the Table 2 that presents the area of the difference of omission between two data sets. The forward-based strategy takes the largest data set (in the example, data set B) and from that, it selects the *non previously selected* data set with respect to which the current data set has the largest difference of omission. This strategy avoids retrieving data sets with no new information. In this case, the sequence is given by B, D, A, and C.

- *History-based strategy:* This strategy uses the information about the relationship of already selected data sets to define which data set is the next in the navigation process. The idea behind this strategy is that all previous data sets, and not only the current data set, may also contain part of the data of a non-previously selected data set. So, new information in the navigation process may be less than the data commission of non-selected data sets with respect to the current data set. Like the forwarded-based strategy, history-based strategy avoids accessing data sets with no new information.

$O()$	A	B	C	D
A	0	**12**	5	11
B	2	0	7	**8**
C	3	**15**	0	8
D	6	**13**	5	0

Table 2. An example of navigation with the forward-based strategy

Estimating the area of expected new regions based on previously selected data sets assumes that all regions are equally likely to be shared among data sets. Therefore, when considering any two data sets, the ratio between the area of equivalent objects and the total area in one of the data sets estimates the *likelihood* of a region in this data set of being in the set of equivalent objects.

To illustrate our derivation of the area of expected new regions in a navigation process, consider a case of a non previously selected data set A with area n and two already selected data sets B and C. The graph of the corresponding information space gives the regions that are equivalence between A and B and between A and C, $|E(A,B)|$ and $|E(A,C)|$, respectively. Likewise, the regions in A that are not in B and C, $|C(A,B)|$ and $|C(A,C)|$, respectively, are known. In this case, we distinguish the following cases when deriving the expected number of potential new regions when visiting A ($NEW(A)$):

1. When one of the previously selected data sets contains all regions stored in data set A (Figure 3), no new regions can be retrieved from visiting A:

$$|E(A,B)| = n \vee |E(A,C)| = n \rightarrow NEW(A) = 0$$

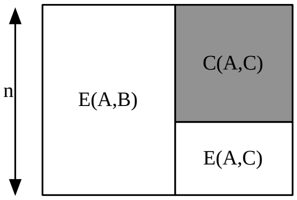


Fig. 3. History-based strategy: a case with no new regions

2. When one of the previously selected data sets contains none of the regions in A, and the other previously selected data set contains a partial set of regions in A (Figure 4), the expected number of new regions when visiting A is equal to the area of regions that *both* previously selected

data sets did not contain.

$$|C(A,B)| \neq n \wedge |C(A,C)| = n \rightarrow NEW(A) = |C(A,B)| \vee$$
$$|C(A,B)| = n \wedge |C(A,C)| \neq n \rightarrow NEW(A) = |C(A,C)|$$

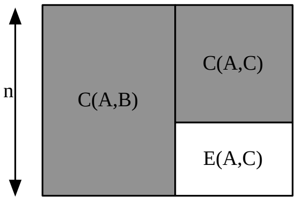


Fig. 4. History-based strategy: a case with new regions

3. When $|E(A,B)| = |C(A,B)|$ and $|E(A,C)| = |C(AC)|$ (Figure 5), the expected area of new regions is equal to the half of $|C(A,B)|$ or $|C(A,C)|$, since we consider the same probability for all regions of being equivalent or different between two data sets. In the general case with two previously selected data sets, the estimated area of new regions is:

$$\begin{aligned} |C(A,B)| \neq n \wedge |C(A,C)| \neq n \rightarrow & \\ NEW(A) &= n \times \left(\frac{|C(A,B)|}{n}\right) \times \left(\frac{|C(A,C)|}{n}\right) \\ &= \frac{|C(A,B)| \times |C(A,C)|}{n} \end{aligned}$$

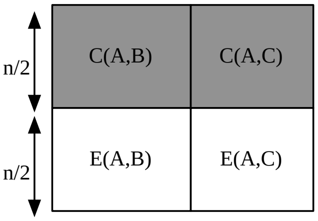


Fig. 5. History-based strategy: a case with half of new objects

The general expression of the expected new regions in a non-selected data set X of area n, when m data sets $\{Y_1, \ldots, Y_m\}$ have been previously visited is:

$$NEW(X) = \frac{\prod_{i \neq j} |C(Y_i, Y_j)|}{n^{m-1}} \tag{10}$$

In the example of Figure 2, this strategy takes the data set with the largest area and continues accessing the non-selected data sets in the following sequence: B, D, C and A.

4.2 Incomplete Information Space

Since the cost of constructing a complete information space may be very high, this section analyzes the strategies in presence of an incomplete representation of the information space. In this analysis, the total area of objects per data set is known, and the data equivalence, difference of data commission, and difference of data omission may be unknown. The focus of the analysis is on the equivalence, since knowing the data equivalence between data sets and the total area per data set allows us to quantify the differences between data sets.

The size-based strategy is not affected by the incomplete information space. The last two strategies, forward- and history-based strategies, however, need some derivation or approximation. Useful to this approximation are the lower and upper bounds defined by composition of correspondence relations described in Section 2.2. Such an approximation requires a connected information space (connected graph); that is, there exists at least one path between any two nodes in the graph representation. Consider for example the connected and incomplete information space in Figure 6.

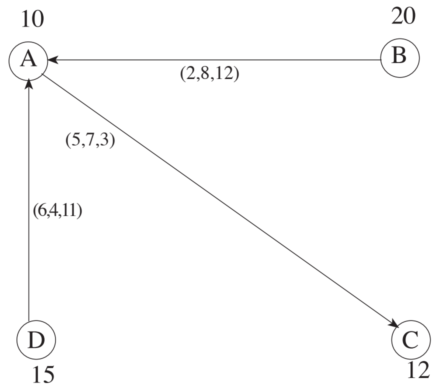


Fig. 6. Example 1: Incomplete and connected information space

In the information space of Figure 6, three edges of correspondence relations are missing: edge between B and C, edge between B and D, and edge between C and D. Based on the Equation 7, the lower and upper bounds of correspondence relations are:

$$\begin{array}{lll}
5 \leq |E(B,C)| \leq 12 & 2 \leq |E(B,D)| \leq 15 & 1 \leq |E(C,D)| \leq 12 \\
8 \leq |C(B,C)| \leq 15 & 5 \leq |C(B,D)| \leq 18 & 0 \leq |C(C,D)| \leq 11 \\
0 \leq |O(B,C)| \leq 7 & 0 \leq |O(B,D)| \leq 13 & 3 \leq |O(C,D)| \leq 14
\end{array}$$

These bounds are used in the navigation strategies to approximate the value of non-previously retrieved regions (new regions) that will be obtained when accessing a non-retrieved data set. The *forward-based* strategy takes the difference of omission ($|O()|$) of the last retrieved data set and each of the non-retrived data sets. In an interval of possible values for $|O()|$, all these values are considered to be equally possible, so the strategy uses the media of the interval.

The *history-based* strategy requires the numbers $|E()|$ and $|C()|$ of non-retrieved data sets with respect to previously retrieved data sets. Like the *forward-based* strategy, the *history-based* strategy takes the media values of the derived intervals that bound $|E()|$ and $|C()|$. The resulting approximations are:

$$\begin{array}{lll}
|E(B,C)| \backsim 8.5 & |E(B,D)| \backsim 8.5 & |E(C,D)| \backsim 6.5 \\
|C(B,C)| \backsim 11.5 & |C(B,D)| \backsim 11.5 & |C(C,D)| \backsim 5.5 \\
|O(B,C)| \backsim 3.5 & |O(B,D)| \backsim 6.5 & |O(C,D)| \backsim 8.5
\end{array}$$

When applying the approximations to the navigation strategies in the example given in Figure 6, the sequence in which data sets are accessed is the one given in Table 3. The results in this Table matches the results for the size-base and history-based strategies of the complete graph.

Order	Size-based	Forward-based	History-based
1	B	B	B
2	D	D	D
3	C	C	C
4	A	A	A

Table 3. Sequence of retrieved data sets for an incomplete information space

In the previous example, all non-existing edges (unknown correspondence relations) were derived by the composition of existing edges (known correspondence relations). In same cases, a derived edge must be needed to derive a second edge. Consider the example in Figure 7.

In this example, to derive the correspondence relations between C and D requires having correspondence relations between A and C or between B and D, both of them being derived relations. In such case, we consider the process of completing the information in two steps: (1) a derivation from only known correspondence relations and (2) a derivation that combines known and unknown categories.

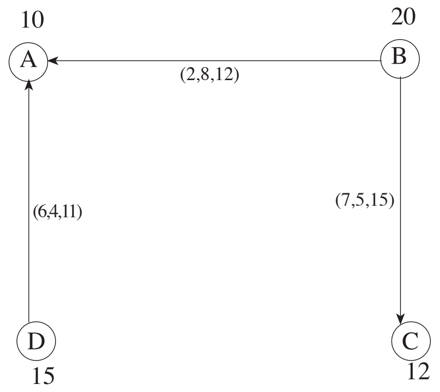


Fig. 7. Example 2: Incomplete and connected information space

An algebraic analysis of using a derived correspondence relations to derive a new correspondence relations follows. Constants are the total areas per data set or the known correspondence relations. Consider the example in Figure 7, the constants a, b, c are the areas in data sets A, B, and C, respectively. The derivation of $|E(A,C)|$ follows the Equation 11.

$$max(0, |E(A,B)| - (b - |E(B,C)|) \leq |E(A,C)| \leq min(a,c) \tag{11}$$

If $d \leq |E(A,B)| \leq e$, where d and e are previously derived lower and upper bounds, then

$$max(0, d - (b - |E(B,C)|) \leq |E(A,C)| \leq min(a,c)$$

includes all possible values of $|E(A,C)|$, since the effect of the upper bound of $|E(A,B)|$ is already considered within the interval between the lower bound of $|E(A,B)|$ and the upper bound of $|E(A,C)|$. Likewise, if $d \leq |E(B,C)| \leq e$, then

$$max(0, |E(A,B)| - (b - d) \leq |E(A,C)| \leq min(a,c)$$

includes all posssible values.

When deriving the correspondence relations, if there exist more than one composition path between data sets, it is the intersection of all intervals derived from composition paths of length 2 between data sets what defines the interval of a correspondence relation. Such intersection comes from the definition of path consistency of logical consistency in a graph [11].

In the intersection of these intervals, upper bounds of different composition paths between two data sets are the same, since these intervals are bounded by the minimum of total objects between data sets. For lower bound, the intersection is equivalent to the maximum of lower bounds. For example, the derivation of $|E(C, D)|$ in Figure 7 can be done by using $|E(C, B)|$ and $|E(B, D)|$ or by using $|E(C, A)|$ and $|E(A, D)|$. Thus, it is the intersection of both results what defines $|E(C, D)|$. The derived intervals of the example in Figure 7 are:

$$\begin{array}{l}
0 \leq |E(A, C)| \leq 10 \; 2 \leq |E(B, D)| \leq 15 \; 0 \leq |E(C, D)| \leq 12 \\
0 \leq |C(A, C)| \leq 10 \; 5 \leq |C(B, D)| \leq 18 \; 0 \leq |C(C, D)| \leq 12 \\
2 \leq |O(A, C)| \leq 12 \; 0 \leq |O(B, D)| \leq 13 \; 3 \leq |O(C, D)| \leq 15
\end{array}$$

With the previously estimated values for the areas of correspondence relations, the sequence in which data sets are accessed is as follows:

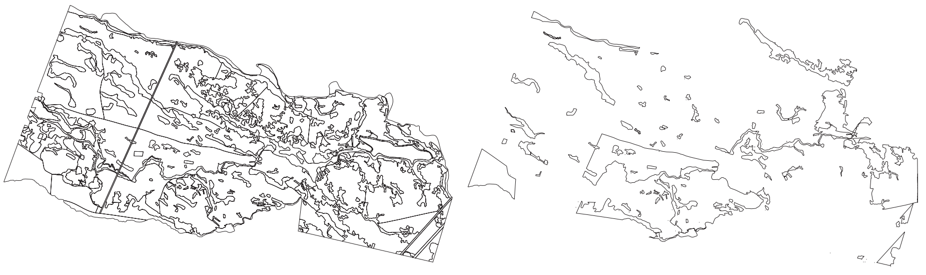
Order	Size-based	Forward-based	History-based
1	B	B	B
2	D	C	C
3	C	D	D
4	A	A	A

Table 4. Sequence of retrieved data sets for an incomplete information space with double derivation

The different configurations of incomplete graphs have an effect on the navigation strategies. Therefore, when handling incomplete graphs, the configurations with less needed derivations should be chosen (i.e., the direct derivation from known correspondence relations). An example is a star-like configuration, with the center being the data set with the largest area of objects.

5 Implementation of Strategies

To illustrate how the strategies for information navigation can be implemented, we use real data taken from the domain of a Forestry Cadastral System. We consider a thematic layer with 406 polygons of protected regions and we created sets of 10 different data sets. These data sets were created by randomly selected a percentage of the objects in the thematic layer. In the following illustrations we only used from 20% to 40% of the objects so that we could visually appreciate the differences between data sets. In addition to selecting different sets of objects, we introduce geometry changes over a percentage of the selected objects (10%). These changes were translations in one or both dimensions equivalent to a random percentage of the size of the objects in the chosen directions. Thus, we expected to consider two cases when matching different data sets: incomplete information and partial overlapping. Figure 8 shows the complete thematic map and one data set derived with the process described above (20% of regions with a 10% of changes).



(a) (b)

Fig. 8. Data: (a) complete thematic layer and (b) a derived data set

To implement the strategies, we approximated objects by their minimum boundary rectangles (MBRs). This approximation creates overlapping areas within a data set. We do so no only for computational simplicity of our implementation, but also because MBRs are the common approximation used by spatial databases for access methods and searching. Our final goal is to use these strategies on-line (i.e., with time constraints) and to relate our navigation strategies with searching in current databases.

We run the tree navigation strategies: size-based, forward-based and history-based strategies. The results were compared by determining the total regions (without duplication) that are actually obtained by accessing one after another data set. For example, consider only 3 data sets A, B, and C with an access sequence equal to B, C and A. The evaluation will say that for the first access we gain all data in B, for the second access, we gain all data in $C \cup B$, and for the last access, we gain all data in $A \cup B \cup C$. This can be represented in a graph in terms of percentages of new data (Figure 9).

The graph in Figure 9 indicates that the history-based strategy retrieves new regions quicker than the other two strategies. The history-based strategy best estimates the information content obtained in the retrieval process. With only accessing the tree first ranked data sets, this strategy obtains over 98% of information content in all data sets. The behavior of the size-based strategy is explained for the fact that two large data sets may have many duplicated regions so that we may not gain much information with accessing both of them. The forward-based strategy is limited to look to the next data set without considering information of the previous to the last data set that was selected.

Figure 10 shows three of the ranked data sets: the first, the second and last ranked data set. It is possible to appreciate that the last ranked data set does not provide much new information.

The results shown here are examples of what we obtained with different sets of data. In all cases, the history-based strategy outperforms the forward-based and size-based strategies. Even more, in many cases, we needed only two o three data sets to obtain all information. Such cases consider data sets with larger

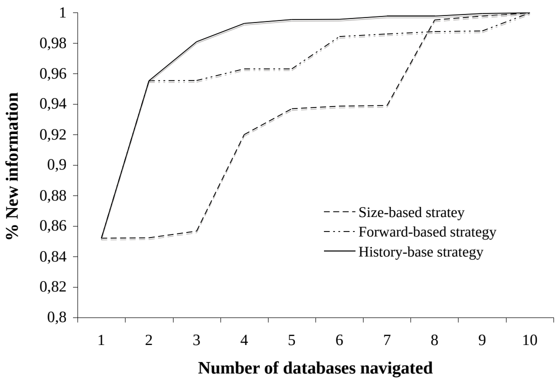
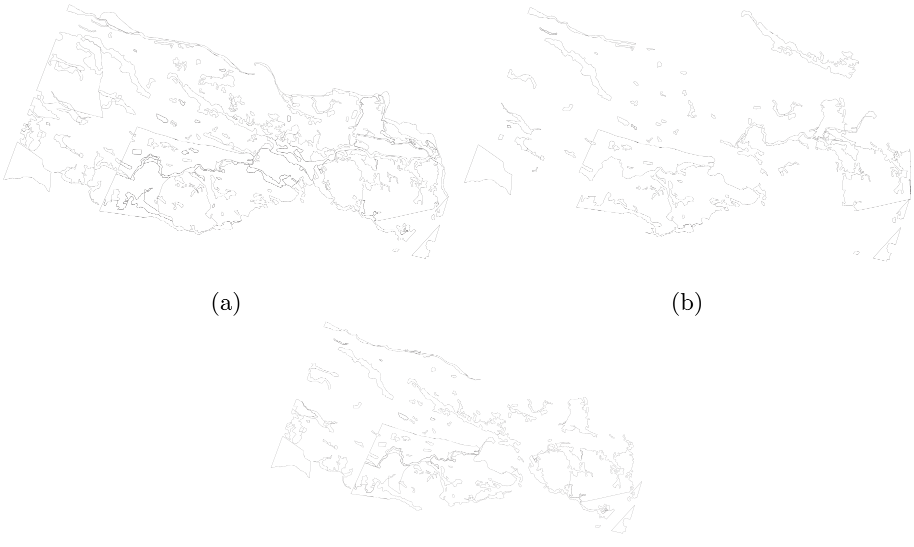


Fig. 9. Results in terms of percentages of new information



(a)

(b)

(c)

Fig. 10. Ranked data sets: (a) first, (b) second, and (c) last data set

numbers of objects than the number of objects in the illustrated example and, therefore, data sets with larger numbers of duplications.

Finally, we considered an incomplete graph and we run the history-based strategy. The incomplete graph had a star-like configuration with one of the largest data sets being in the center of the star. The results in terms of the percentages of new information are shown in the graph of Figure 11. This graph shows no important differences between the results of the strategy with complete and incomplete graphs.

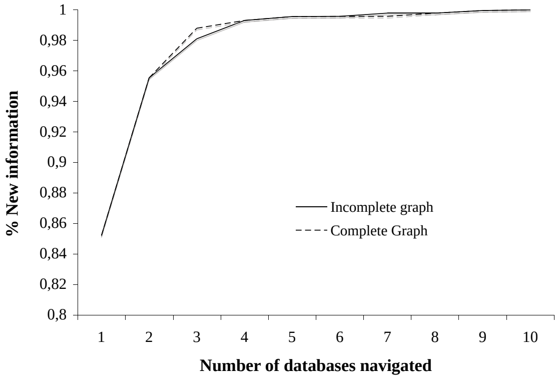


Fig. 11. Results for a complete and an incomplete graph

6 Conclusions

This work describes a formalism for relating the information content of data sets. Based on this characterization, it proposed a way to navigate or access these data sets to optimize the retrieval process. The experimental results show promising results for being applying in real cases of heterogeneous databases.

There are interesting topics for further investigation. We will make further evaluations of the strategies by comparing what the real impact of using MBR on the navigation strategies is. In this work we use the area of regions, but we expect to match objects and use the number of equivalent versus different objects in the navigation strategies. We would like to compare these two approaches: region versus objects. An additional evaluation will analyze the different strategies under incomplete information.

Using this formalism for query processing implies that one needs to characterize the information content of data sets associated with query windows in spatial databases. Since an on-line characterization seems inefficient and a characterization at the whole databases does not provide enough information of a particular spatial window, pre-defined aggregations at different levels of space partitions are needed. These aggregations may follow the organization given for spatial access methods and may use methods that approximate correspondence relations of overlapping regions.

Finally, we are planning to extend our work by assuming that additional information may exist about the origin of the data. In such situation, some databases may be considered more reliable than others, and some kind of database ranking may be included into the navigation strategies.

Acknowledgment. Andrea Rodríguez's research work was funded by FONDECYT 1050944, CONICYT, Chile. Francisco Godoy is funded by FONDAP-COPAS 150100007, CONICYT, Chile.

References

1. ACM. http://www.acm.org/class. 2006.
2. R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval.* ACM Press, 1999.
3. D. Cutting, D. Karger, and J. Oederson. Constant interaction-time scatter/gather browsing of very large document collection. In *16th Annual International ACM/SIGIR Conference*, pages 126–135, 1993.
4. M. Egenhofer, E. Clementini, and P. Di Felice. Evaluating inconsistency among multiple representations. In *Spatial Data Handling*, pages 901–920, Edinburg, Scotland, 1994.
5. M. Egenhofer and J. Sharma. Assessing the consistency of complete and incomplete topological information. *Geographical Systems*, 1(1):47–68, 1993.
6. R. Flowerdew. *Spatial Data Integration*, chapter Spatial Data Integration, pages 375–387. Longman Scientific & Technical, 1991.
7. F. Fonseca, M. Egenhofer, P. Agouris, and C. Camara. Using ontologies for integrated information systems. *Transactions in GIS*, 6(3):231–257, 2002.
8. A. Rodríguez and M. Egenhofer. Determining semantic similarity among entity classes from different ontologies. *IEEE Transactions on Knowledge and Data Engineering*, 15(2):442–456, 2003.
9. V. Kashyap and A. Sheth. Schematic and semantic similarities between batabase objects: A context-based approach. *The Very Large Database Journal*, 5(4):276–304, 1996.
10. B. Kuipers, J. Paredaens, and J. den Busshe. On topological equivalence of spatial databases. In F. Afrati and Ph. Kolaitis, editors, *6th International Conference on Database Theory ICDT97, LNCS 1186*, pages 432–446. Springer Verlag, 1997.
11. A. Mackworth. Consistency in networks of relations. *Artificial Intelligence*, 8(1):99–118, 1977.
12. E. Mena and A. Illarramendi. *Ontology-Based Query Processing for Global Information Systems.* Kluwer Academic Publishers, Norwell, MA, 2001.
13. S. Ram and S. G. Modeling and navigation of large information spaces: A semantic based approach. In *International Conference on System Science.* [http://computer.org/proceedings/hicss/0001/00016/00016020abs.htm], IEEE CS Press, 1999.

14. D. Roussinov and M. McQuaid. Information navigation by clustering and summary query results. In *International Conference on System Sciences*, page 3006. IEEE CS Press, 2000.
15. A. Sheth. *Interoperating Geographic Information Systems*, chapter Changing Focus on Interoperability in Information Systems: From System, Syntax, Structure to Semantics, pages 5–30. Kluwer Academic Plublishers, 1999.
16. VIVISMO. http://vivismo.com. 2006.
17. P. Weinstein and P. Birmingham. Comparing concepts in differentiated ontologies. In *12th Workshop on Knowledge Adquisition, Modeling, and Management*, Banff, Canada, 1999.
18. M. Worboys and E. Clementini. Integration of imperfect spatial information. *Journal of Visual Languages and Computing*, 12:61–80, 2001.
19. Yahoo! http://www.yahoo.com. 2006.
20. O. Zamir and O. Etzioni. Grouper: A dynamic cluster interface to web search results. In *WWW8*, 1999.