Supporting Visual Information Extraction from Geospatial Data

Giuseppe Della Penna and Sergio Orefice
Department of Computer Science
University of L’Aquila, Italy
Email: {giuseppe.dellapenna,sergio.orefice}@univaq.it

Daniele Magazzeni
Department of Sciences
University of Chieti-Pescara, Italy
Email: magazzeni@sci.unich.it

Abstract—The Spatial Relation Query (SRQ) tool is a graphical software system, supported by a SQL-like query language, that enables users to perform information extraction driven by the visual appearance and the spatial arrangement of the information. The tool has been initially designed to support visual information extraction from web pages. Indeed, its former underlying spatial relation formalism relied on the bounding boxes of the graphical objects, which is a suitable choice for the web domain. In this paper we present a theoretical extension of the SRQ spatial composition framework that has been enhanced to work directly on the contours of the graphical objects. This allows us to apply the tool to more general contexts, such as GIS applications.

I. INTRODUCTION

The process of knowledge acquisition from generic information domains has its central phase in the Information Extraction (IE), which aims to extract from the located documents relevant information that appear in certain semantic or syntactic relationships. In particular, IE tries to process the relevant information found on the documents in order to make it available to structured queries. Most often, information extraction systems are customized for specific application domains, and require manual or semi-automatic training sessions.

In [1] we have proposed a general IE approach based on the visual appearance of the information, conceived as its user-perceived rendering. This allows one to shift the IE problem from the low level of code (e.g., raster graphics, vector drawing, wordprocessor formatted text, web page, etc.) to the higher level of visual features, providing a paradigm of the kind “what you see drives your search” that supports a natural query formulation. The approach is based on the box spatial relation theory [2], such that graphical objects are syntactically described and manipulated through their bounding boxes whatever the shape of the object is, and on a SQL-like language, which allows users to write queries based on the visual arrangement of the information in an intuitive way. These formalisms have been implemented in a full-featured graphical software system, the SRQ (Spatial Relation Query) tool, which has been profitably used on a wide variety of applications within the IE web page domain. An early attempt to apply the tool to geospatial data has been proposed in [3]. However, in that work the underlying box spatial relation theory could deal with only very simple applications.

Other approaches in the recent literature that make visual information extraction include, for example, techniques that focus on specific application areas like, e.g., the work in [4] where the authors propose a machine learning methodology which allows one to automatically extract specific field of PDF documents, or the approaches based on visual web page analysis [5]–[8] which exploit the visual web page representation. However, the latter works focus on information extraction specifically targeted on tasks like record boundary detection [5], web page segmentation [7], visual web table extraction [6], or visual similarities detection (e.g., the recognition of repetitive patterns) [8]. Moreover, the major shortcoming of all the approaches above exploiting the visual appearance of information is the lack of an automatic counterpart supporting the visual information extraction.

In the last years, several studies have focused on the adoption of visual languages and visual techniques in many critical activities related to Geographic Information Systems (GIS) [9], ranging from design of geographic databases, to interoperability support, to decision-making support (see, e.g. [10]–[13]). Motivated by the increasing attention of researches aiming at bridging the gap between advanced geographic management techniques and practical problem-solving, in this paper we present the extension of the SRQ framework proposed in [1], [3] to the GIS context. To this aim, the spatial relation theory has been reformulated in order to work on polygonal contours. Indeed, the previous bounding box syntactical model has some intrinsic limitations that make it difficult to apply in frameworks where graphical objects are represented by complex figures, such as in GIS applications. The new theory has been embedded in the SRQ tool, which now is able to be experimented also on complex and meaningful applications in the geospatial data domain.

II. THE SPATIAL COMPOSITION FRAMEWORK

Before describing the new spatial relation formalism, let us briefly recall the basic notions of the box syntactical model as defined in [2]. In general, a syntactical model describes a family of visual languages based upon the nature of their graphical objects and composition rules, and it is formally defined by the quadruple (graphical image, syntactical image, syntactical attributes, spatial relations), where the first three components characterize the graphical objects by means of
their graphical (namely the graphical image) and logical (namely the syntactical image and attributes) parts. In particular, the syntactical image is a suitable approximation of the graphical image that makes easier its syntactical manipulation, whereas the syntactical attributes are specific points on the syntactical image used by the spatial relations to compose graphical objects and form visual sentences.

Formally, in the box syntactical model, graphical objects have general geometric figures as their graphical images, and their bounding boxes as the corresponding syntactical images. The syntactical attributes of a graphical object are the upper left and lower right points of the bounding box, and the spatial relations are specified as follows.

**Definition 1:** Given a graphical object \( a \) with syntactical attributes \((x, y)\) and \((x_1, y_1)\), a generic spatial relation \( REL \), with respect to \( a \), defines two functions \( UL_{REL}(x, y, x_1, y_1) \) and \( LR_{REL}(x, y, x_1, y_1) \) that map the coordinates of the upper-left \((x, y)\) and lower-right \((x_1, y_1)\) points of the bounding box of \( a \) onto sets of points.

Then, given a graphical object \( b \) with syntactical attributes \((h, k)\) and \((h_1, k_1)\), we have that \( aRELb \) holds if and only if \((h, k) \in UL_{REL}(x, y, x_1, y_1)\) and \((h_1, k_1) \in LR_{REL}(x, y, x_1, y_1)\).

In other words, \( aRELb \) holds if and only if the upper left and the lower right points of the bounding box of \( b \) are contained respectively within two areas calculated on upper left and lower right points of the bounding box of \( a \) through the functions \( UL_{REL} \) and \( LR_{REL} \) (see Figure 1).

Definition 1 yields three types of possible spatial arrangements: inclusion, intersection and spatial concatenation, where the term “concatenation” refers to any spatial arrangement of graphical objects not intersecting their areas. As an example, the relations \( INTERSECT, INCLUDE, UP, DOWN, LEFT \) and \( RIGHT \), taken from [2], are defined in Table I. These relations model the general types of spatial overlapping and spatial concatenation.

So far, the box syntactical model has been used as the underlying formalism of the SRQ tool to support visual information extraction from web pages in [1] and from simple maps in [3]. However, this model has some intrinsic limitations, since the graphical objects are syntactically described and manipulated through their bounding box, whatever the shape of the object is. As a consequence, it is difficult to apply it in frameworks where graphical objects are represented by complex figures, such as in general GIS applications. Indeed, in the box syntactical model, the syntactical image (i.e., the bounding box) of a graphical object matches its graphical image (i.e., its real shape), only for rectangular figures, whereas the bounding box for any other kind of object typically includes areas that are not actually part of the object itself (see, e.g., the objects in Figure 2). Due to this fact, the application of spatial relations may sometimes lead to results that are actually incorrect. For example, Figure 2 shows the typical image intersection problem, where the spatial relation \( INTERSECT \) would hold between the two (disjoint) shapes, since their bounding boxes actually intersect. To overcome this limitation, it is necessary to conceive the syntactical image of a graphical object as its graphical image. This idea is at the basis of the extension to new contour syntactical model.

**A. The Contour Syntactical Model**

The graphical objects of the visual languages characterized by the new model have “simple polygons” as their graphical images. In geometry, a polygon is a region of the plane bounded by a finite collection of line segments forming a closed curve. A simple polygon is a polygon where no pair of edges cross each other.

Figure 3(a) shows a simple polygon, whereas the contour in Figure 3(b) is an example of non simple polygon since it contains self loops.

Moreover, in order to support a more precise manipulation,
the logical part of graphical objects has to be appropriately extended. Thus, the other novelty of the contour syntactical model is in the syntactical image component of the graphical objects, which now coincides exactly with the contour of the graphical image.

Then, a natural choice for defining the graphical object syntactical attributes in the new model is to consider the smallest number of points of the syntactical image that allow one to completely enclose the graphical object (as it was also implicitly done in the box syntactical model with the upper left and lower right points).

Formally speaking, given a graphical object $O = (G, S, A, R)$ with syntactical image $S$, we write $p \in S$ to indicate a generic point of $S$ and $p_x, p_y$ to indicate the $x$ and $y$-coordinate, respectively, of $p$. Moreover, we shall indicate with $UP(S)$, $DP(S)$, $LP(S)$ and $RP(S)$ the upmost, downmost, leftmost and rightmost point, respectively, of $S$. For instance, $UP(S)$ is formally defined as $p \in S | \forall p' \in S, p'_y \leq p_y$.

In particular, in the contour syntactical model, the syntactical attributes $A$ of a graphical object with syntactical image $S$ are the set of four points $UP(S)$, $DP(S)$, $LP(S)$ and $RP(S)$, when used by spatial concatenation relations, or the set of all the points of the contour $S$, in the case of inclusion and intersection relations. The corresponding setting of spatial relations is formally defined in the following subsection.

B. Contour Spatial Relations

The new characterization of the logical part of graphical objects in the contour model leads to the following reformulation of the spatial relations.

**Definition 2:** A generic spatial relation $REL$ between two graphical objects $a$ and $b$, with syntactical attributes $A$ and $A'$, respectively, is defined by means of the function $set_{REL}(A,p) = X \in P(R^2)$ with $p \in A'$ that associates each point in $A'$ to a set of points $X$. Then, a $REL$ relation holds iff $\forall p \in A', p \in set_{REL}(A,p)$.

In other words, each syntactical attribute of $b$ must be contained in the corresponding set of points $X$ calculated by the syntactical attributes of $a$, as appropriately defined by the function $set_{REL}$.

Let $a$ and $b$ be two graphical objects with syntactical image $S$ and $S'$ and syntactical attributes $A$ and $A'$, respectively. Then, in order to define the four basic concatenation spatial relations (which model the disjoint spatial arrangement) in the contour model, it is enough to consider $A = \{UP(S), DP(S), LP(S), RP(S)\}$ and $A' = \{UP(S'), DP(S'), LP(S'), RP(S')\}$, as shown in Table II. Note that, to shorten the notation, in the table we assume that $set_{REL}$ returns $R^2$ if not differently specified.

For example, Figure 4 shows a visual arrangement where the $LEFT$ relation holds between the objects $a$ and $b$.

To complete the set of basic contour spatial relations, let us define the inclusion and intersection relations that model the overlapping of graphical objects. Let $a$ and $b$ be two graphical objects with syntactical image $S$ and $S'$ and syntactical attributes $A$ and $A'$, respectively. In order to define the two basic relations $INCLUDE$ and $INTERSECT$ in the contour model, we have to consider $A = S$ and $A' = S'$. Then, we define such relations as shown in Table III.

These two relations use the auxiliary relation $CONTAIN(p)$, defined through the function $set_{CONTAIN}(A,p) = Area(S)$, where $Area(S)$ is the set of internal points of the syntactical image $S$. In other words, $a \text{ } CONTAIN(p) \text{ } b$ iff the point $p \in S'$ is inside the area of $S$.

Thus, a graphical object $a$ includes another graphical object $b$ iff all the points of the contour of $b$ are enclosed by the contour of $a$. On the other hand, a graphical object $a$ intersects another graphical object $b$ iff at least one of the points of the contour of $b$ is enclosed by the contour of $a$.

III. THE SRQ TOOL

The overall organization of the SRQ software system is shown in Figure 5. The current prototype application has been developed in Java with the support of ANTLR [14] for the spatial relation compiler and query interpreter. The application

### Table II

**Definition of basic concatenation contour spatial relations**

<table>
<thead>
<tr>
<th>Relation</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>UP</td>
<td>$set_{UP}(A, UP(S')) = { (m, n)</td>
</tr>
<tr>
<td>DOWN</td>
<td>$set_{DOWN}(A, DP(S')) = { (m, n)</td>
</tr>
<tr>
<td>LEFT</td>
<td>$set_{LEFT}(A, LP(S')) = { (m, n)</td>
</tr>
<tr>
<td>RIGHT</td>
<td>$set_{RIGHT}(A, RP(S')) = { (m, n)</td>
</tr>
</tbody>
</table>

### Table III

**Definition of the overlapping contour spatial relations**

<table>
<thead>
<tr>
<th>Relation</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a \text{ INCLUDE b}$</td>
<td>$\forall p \in S : a \text{ CONTAIN}(p) \text{ b}$</td>
</tr>
<tr>
<td>$a \text{ INTERSECT b}$</td>
<td>$\exists p \in S : a \text{ CONTAIN}(p) \text{ b}$</td>
</tr>
</tbody>
</table>
uses a plugin system to interface with several different sources and build the corresponding visual data model.

In particular, the SRQ tool is composed of two interacting modules: the Spatial Relation Editor and the Spatial Relation Query Executor. These modules share the access to a spatial relation library which contains the basic building blocks of the query language.

The Spatial Relation Editor module allows one to create and manage the spatial relation library, which defines the spatial relations used in the queries. The library itself is written in an appropriate Spatial Relation Definition Language (SRDL) that, for sake of simplicity, will not be described here.

The Spatial Relation Query Executor allows one to compose and execute queries. The composition of queries can be accomplished through the Spatial Relation Query Language (SRQL) language, which exploits the spatial relation theory given in Sections II-A, II-B to create queries based on the visual arrangement of the data and is designed to be as similar as possible to the well-known SQL.

Before applying an SRQL query, the tool analyzes the information resource and generates the corresponding user image, i.e., an image which shows how the information actually looks like to the user. Then, it builds an appropriate visual data model, namely the contour model, that implements the notion of graphical object as the 4-tuple formally defined in Section II-A. This model encapsulates all the distinguished objects from the visual appearance, possibly associating a set of attributes, deducible from the source information, to each object. Depending on the particular IE data domain, such attributes characterize specific properties of the objects, which may be also visual (e.g., color, contained text, etc.).

The basic structure of a SRQL query is shown in Figure 6. More details on the syntax of the clauses composing a query can be found in [3].

Once a query has been written, it can be saved for later reuse, and analyzed/executed. Indeed, the system first performs some syntax checks on the query, possibly providing meaningful error messages that allow the user to improve the query definition, and then the query is executed on the contour model. After a successful query, the user image of the data is rendered in the application interface, the graphical objects selected by the query (if any) are highlighted, and the corresponding attributes are appropriately returned. Finally, it is worth noting that the query results can be exported in an XML file to be exploited for further analysis and manipulations.

IV. A CASE STUDY

In this section we describe an application of the SRQ tool on real-world GIS data concerning the malaria diffusion. Malaria is an infectious disease widespread in tropical and subtropical regions, including parts of the Americas, Asia, and Africa. Each year, in these regions, malaria kills millions of people.

It is well known that the distribution of this disease is closely related to climatic and environmental factors like temperature, humidity and rainfall. Indeed, malaria is naturally transmitted by the bite of a female Anopheles mosquito which lives in wet and warm places.

First of all, we use SRQ to highlight such a relation by combining and analyzing climate data to find which nations are potentially at risk of malaria. The results of this analysis are then validated against the map shown in Figure 9(b), which illustrates the actual diffusion of the disease as reported by the World Health Organization [15].

In particular, the data sources for this case study are represented by the three maps taken from FAO [16] which contain, respectively, the average humidity, annual rainfall and temperature in the world. As an example, Figure 7(a) shows the temperature map. We use the MapWindow [17] software to trace such maps and build the corresponding shapefiles. Figure 7(b) shows the temperature shapefile, where we report the most significant map areas, associating each of them with the semantic data given by the corresponding legend.

Finally, these shapefiles have been layered over the world nations shapefile to obtain the composite map to be used as the source for the query to find the nations at risk of malaria.
SELECT *
FROM local:maps
WHERE CONTOUR INTERSECT ANY rainyArea
HAVING layer = 'world'
WITH rainyArea = SELECT CONTOUR
FROM local:maps
WHERE CONTOUR INTERSECT ANY warmArea
HAVING layer = 'rain' AND rain_mm >= 1500
WITH warmArea = SELECT CONTOUR
FROM local:maps
WHERE CONTOUR INTERSECT ANY humidArea
HAVING layer = 'temperature' AND
    degrees_avg >= 20
WITH humidArea = SELECT CONTOUR
FROM local:maps
HAVING layer = 'humidity' AND rate = 5

Fig. 5. SRQ tool architecture

Fig. 8. SRQL query to find the world nations at risk of malaria

which is formulated in the SRQL language as illustrated in Figure 8.

The query uses the INTERSECT relation to intersect the world nations with the areas, taken from the corresponding layers, that represent rainy (i.e., average annual rain greater than or equal to 1500mm), warm (average temperature above 20°C) and humid (humidity index 5) regions. In particular, each of these classes of regions is extracted by a subquery which works on the corresponding map layer.

The query results are shown in Figure 9(a), where the extracted areas are highlighted in yellow. Such results are very similar to the WHO data in Figure 9(b). The only visible difference is that the SRQ ones include more regions of the United States than those marked in the WHO map. However, we should expect this discrepancy, since the southern regions of the United States would be actually exposed to malaria, but the social and economic conditions of such regions (which were not considered by the query) allowed to eradicate the disease through an adequate health care and environmental control.

Encouraged by the positive validation of the SRQ analysis with respect to the current climatic data, we can try to obtain a forecast of the future diffusion of malaria. Indeed, the well known global warming is constantly changing the climatic conditions of many world areas, and this may affect the
incidence of malaria. This forecast actually represents the valuable contribution of SRQ for this case study.

To show what may happen, we have taken a map published by the UK’s National Weather Service [18], which reports the estimated temperature growth in the future. We combined the data from this map with the ones in the temperatures shapefile (Figure 7(b)), to obtain a new shapefile shown in Figure 10, which reports the expected future average temperatures in the world. Finally, we have run again the query in Figure 8 using this modified data together with the same “rain” and “humidity” layers.

The query results, highlighted in yellow in Figure 11, indicate that the malaria could appear in the future in some areas of the southern Europe as well as in Canada and Australia. Interestingly, there have been actually some cases of malaria reported in the last years in Canada [19], [20] and Australia [21].

V. CONCLUSIONS

In this paper we have presented an extension of the bounding box spatial relation formalism in order to work directly on polygonal contours. The new theory has been also implemented within the SRQ tool, a graphical software system that provides an automatic support to the visual information extraction from different sources. In particular, the SRQ tool is suitable to be used both by novel or experienced users. The former users may take advantage of the user-friendly interface and the natural query paradigm of the tool to accomplish their searches. The latter ones may fully exploit the integrated power of the tool and of the SRQL language by both appropriately customizing the spatial relations and extensively using visual, textual and structural constraints in the queries.

So far, the SRQ tool has been mainly applied to web case studies. Thanks to the new contour spatial relation formalism, we are experimenting the extended version of the tool on geospatial applications. Other than extensively using the tool in GIS applications, in the future we also plan to apply it to new, challenging domains like, e.g., biomedical applications. Indeed, the contour model is suitable to correctly represent the complex spatial relations existing among the human body organs, and then the SRQ tool may be of help for diagnostic analysis as well as for medicine and surgery didactic activities.

Moreover, we are investigating some extensions of the spatial relation theory in order to capture further aspects of particular visual data sources, e.g., the notion of time. For example, it could be interesting to make queries that address the evolution of GIS datasets, i.e., to reason about spatio-temporal data.

REFERENCES