
RESEARCH IN COMPUTING SCIENCE
ISSN: 1870-4069
Special issue:
Semantics in
Geoinformatics Applications
Marco Moreno-Ibarra
Serguei Levachkine
(Eds.)
Vol. 25
RCS
Research in Computing Science

Special Issue: Semantics in Geoinformatics Applications

Research in Computing Science

Research in Computing Science es una publicación trimestral, de circulación internacional, editada por el Centro de Investigación en Computación del IPN, para dar a conocer los avances de investigación científica y desarrollo tecnológico de la comunidad científica internacional. **Volumen 25**, Noviembre, 2006. Tiraje: 500 ejemplares. *Certificado de Reserva de Derechos al Uso Exclusivo del Título* No. 04-2004-062613250000-102, expedido por el Instituto Nacional de Derecho de Autor. *Certificado de Licitud de Título* No. 12897, *Certificado de licitud de Contenido* No. 10470, expedidos por la Comisión Calificadora de Publicaciones y Revistas Ilustradas. Portada "El abrazo", Óleo sobre tela (140 x 200 cm.), 1980, Jorge González Camarena. El contenido de los artículos es responsabilidad exclusiva de sus respectivos autores. Queda prohibida la reproducción total o parcial, por cualquier medio, sin el permiso expreso del editor, excepto para uso personal o de estudio haciendo cita explícita en la primera página de cada documento. Impreso en la Ciudad de México, en los Talleres Gráficos del IPN – Dirección de Publicaciones, Tres Guerras 27, Centro Histórico, México, D.F. Distribuida por el Centro de Investigación en Computación, Av. Juan de Dios Bátiz S/N, Esq. Av. Miguel Othón de Mendizábal, Col. Nueva Industrial Vallejo, C.P. 07738, México, D.F. Tel. 57 29 60 00, ext. 56571.

Editor Responsable: *Juan Humberto Sossa Azuela*, RFC *SOAJ560723*

***Research in Computing Science** is published by the Center for Computing Research of IPN. **Volume 25**, November, 2006. Printing 500. The authors are responsible for the contents of their articles. All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, without prior permission of Centre for Computing Research. Printed in Mexico City, November, 2006, in the IPN Graphic Workshop – Publication Office.*

Volume 25
Volumen 25

Special Issue: Semantics in Geoinformatics Applications

Volume Editors:
Editores del Volumen

Marco Moreno-Ibarra
Serguei Levachkine

Instituto Politécnico Nacional
Centro de Investigación en Computación
México 2006

ISSN: 1870-4069

Instituto Politécnico Nacional (IPN)
Centro de Investigación en Computación (CIC)
Av. Juan de Dios Bátiz s/n esq. M. Othón de Mendizábal
Unidad Profesional "Adolfo López Mateos", Zacatenco
07738, México D.F., México

http://www.ipn.mx
http://www.cic.ipn.mx

Indexed in LATINDEX
Indexada en LATINDEX

Printing: 500
Impresiones: 500

Printed in Mexico
Impreso en México

Preface

CONGEO 2006 was the Second International Conference on Geoinformatics Applications. It was held in Mexico City, November 21-24, 2006.

This series of biannual conferences (even years), which are satellite events of the biannual (odd years) International Conferences on Geospatial Semantics (GeoS). While GeoS aim at providing a timely forum for the exchange of state-of-the art research results in the areas of modeling and processing of geospatial semantics, CONGEO focuses on the applications of geospatial semantics to next-generation spatial databases and geographic information systems (GIS) as well as personalized geospatial web services. This conference brought together specialists, researchers, engineers, and practitioners throughout the world whose expertise addressed such issues as: GIS applications development; Conceptual, extensional, ontological, and intentional models of geospatial domains; Intelligent tools for the management, analysis, and synthesis of geospatial information; Development of ontology-driven spatial and spatio-temporal databases; Intelligent systems for geospatial data conversion; Semantic processing of geospatial visual information; Urban planning and modeling; Ontology-driven geographic information systems; Interoperability among different geospatial databases; Personalized geospatial web services; Web-mapping applications; Geospatial data mining; Geospatial knowledge representation and processing over different cultures and languages.

This volume contains 9 carefully selected, peer-reviewed, and revised original papers on Geoinformatics applications. The papers are organized in the following sections:

- Geospatial Services and Information Retrieval,
- Geospatial Data Processing,
- Geospatial Analysis and Its Applications.

We hope that the volume will be useful for researches and students working in the corresponding areas of Geoinformatics as well as for all readers interested in Geographic Information Science.

We are gratefully acknowledge the support provided by the sponsors of CONGEO 2006: IPN (Instituto Politécnico Nacional, Mexico), CIC (Centro de Investigación en Computación, Mexico), CONACYT (Consejo Nacional de Ciencia y Tecnología, Mexico). Our special thanks to Ignacio García Araoz, and the members of Geoprocessing Laboratory of CIC for their noble support.

We hope you will enjoy the works reported in this issue and we are looking forward to see you in the next editions of CONGEO in CIC-IPN, Mexico City.

November, 2006

Marco Moreno-Ibarra
Serguei Levachkine

Table of Contents
Índice

Page/Pág.

Geospatial Services and Information Retrieval

Tightly Coupled Geospatial Interfaces for Collaborative Systems and Just-in-Time Information Retrieval Agents

Arno Scharl

Know-Center and Graz University of Technology
Knowledge Management Institute; Inffeldgasse 21a, 8010 Graz, Austria
scharl@tugraz.at

Abstract. Recent advances in collaborative Web technology are governed by strong network effects and the harnessing of collective intelligence through customer-self service and algorithmic data management. As a result, information spreads rapidly across Web sites, blogs, Wiki applications, and direct communication channels between members of online communities who utilize these services. These various media can be integrated by means of ontology-based tools for building and maintaining contextualized information spaces. The complexity of these spaces calls for new interface technologies that enable users to switch between semantic and geospatial topologies with ease. This paper reviews the literature and outlines the application of geospatial technology for building collaborative systems, and for presenting the results of just-in-time information retrieval agents in an intuitive manner. It introduces *Knowledge Planets* as a new interface metaphor that leverages the new generation of geobrowsing platforms such as NASA World Wind and Google Earth as a front-end for semantic services (see http://www.idiom.at/).

Keywords: Tightly Coupled Interface, Geospatial Projection, Collaborative System, Knowledge Planet, Information Retrieval, Annotation Service

1 Introduction

Integrating cartographic data with geo-tagged knowledge repositories, the emerging Geospatial Web "may ultimately be the big disruptive innovation of the coming decade" [1, xxv]. As such, it will catalyze change and enable a broad range of as yet unforeseen applications. The following hypothetical scenario outlines how the envisioned integration of contextualized information spaces and geospatial technology could radically change individual working environments in the not too distant future, impact the workflow within and across organizations, and enrich the interaction between content providers and their target audience:

Kathryn O'Reilly is a knowledge worker who sells her ability to gather, filter and prioritize electronic content. In a virtual world built on contextualized information spaces, Kathryn seamlessly switches between geographic and semantic topologies. She begins her typical working day floating in the virtual space above earth, ready to

Semantics in Geoinformatics Applications
Research in Computing Science 25, 2006, pp. 3-18

navigate the globe and semantic structures via subtle movements of her eyes, and to access an extensive portfolio of add-on functionality through haptic devices. From her elevated position, Kathryn not only observes the rise and decay of topics, but also the unfolding of social structures based on the unique social networks of her friends and business contacts. Across these networks she builds and shares her knowledge repository, and composes media products that are continuously being validated and enriched by the latest news feeds and third-party multimodal sources.

The underlying content management system automatically tailors the format of her articles to the preferences of her regular readers. Kathryn adds, selects, categorizes, aggregates, filters and extrapolates information along multiple dimensions, with minimal cognitive requirements. She can structure her daily workflows, access archives of historic textual and multimedia data, and customize her virtual environment with various communication services to interact with pre-defined or dynamically assembled groups of like-minded individuals. At any point in time, Kathryn may use portions of the information space to initialize what-if scenarios and advanced economic or scientific simulations, investigating the complex interplay between computer-generated and real-world participants.

2 Annotation Services

The unprecedented success of the Web 2.0 and geo-browsing platforms has popularized the process of "annotating the Planet" [2]. Both semantic and collaborative technologies unfold their full potential through network effects and benefit from a critical mass of annotations [3]. At present, however, many metadata initiatives still suffer from the chicken and egg problem of wishing that existing content was retrofitted with metadata [4]. This "capture bottleneck" results from the beneficiaries' lack of motivation to devote the necessary resources for reaching a critical mass of metadata [5]. Manual annotation remains difficult, time consuming and expensive [6]. Automated document enrichment [3, 7] addresses this shortcoming by parsing existing Web resources and annotating content fragments along multiple dimensions.

2.1 Geospatial Annotation

Once geospatial context information becomes widely available, any point in space will be linked to a universe of commentary on its environmental, historical and cultural context, to related community events and activities, and to personal stories and preferences. There are several sources of geospatial context information for annotating knowledge repositories:

- Annotation by the authors, manually [8] or through location-aware devices such as GPS navigation systems, RFID-tagged products, and cellular handsets. These devices geo-tag information automatically when it is being created.

- Determining the location of the server – e.g. by querying the Whois[1] database for domain registrations, monitoring how Internet traffic is routed, or by analyzing the URL for additional cues [4].
- Automated annotation of existing documents. The processes of recognizing geographic context and assigning spatial coordinates are commonly referred to as *geo-parsing* and *geo-coding*, respectively [4].

2.1 Extraction and Disambiguation of Geospatial Context

Electronic resources contain metadata as explicit or implicit geographic references. This includes references to physical features of the Earth's surface such as forests, lakes, rivers and mountains, and references to objects of the human-made environment such as cities, countries, roads and buildings [9]. Addresses, postal codes, telephone numbers, and descriptions of landmarks also allow to pinpoint exact locations [4, 10].

At least 20 percent of Web documents contain easily recognizable and unambiguous geographic identifiers [11]. News articles are particularly rich in such identifiers, since they generally report on the location where an event took place, or where it was reported from [12]. The BBC article "Vienna Marking Mozart Milestone" [13], for example, has a target geography of EUROPE/AUSTRIA/VIENNA, and a source geography of EUROPE/UNITED KINGDOM/LONDON. In addition to target and source geography [14], natural language processing can also be used to extract the geographic scope (= intended reach) of Web resources [15].

Named entity recognition locates and interprets phrasal units such as the names of people, organizations, and places [16, 17]. As with most named entity recognition tasks, ambiguity, synonymy and changes in terminology over time complicate the geo-parsing of Web documents [14, 18, 19]. The heterogeneity of spatial references remains a challenge for even the most sophisticated algorithms. Identical lexical forms may refer to distinct places with the same name (VIENNA referring to the capital of Austria as well as a town in Northern Virginia, US), or can have both geographic and non-geographic meanings. TURKEY denotes both a large gallinaceous bird and a bi-continental country between Asia and Europe. Geo-parsing services need to correctly process references to identical or similar places that may be known under different names, may belong to different levels of the administrative or topographical hierarchy, or may be nearby by some measure of proximity [9].

2.1 Assigning Geospatial Coordinates

Once a location has been identified, the documents can be assigned precise spatial coordinates – latitude, longitude, and altitude – by querying structured geographic indices (gazetteers) for matching entries [20, 21]. This process of associating documents with a formal model is also referred to as "document enrichment" [5, 7]. Examples of formal models in the geographic domain are the Geographic Names Infor-

[1] http://www.whois.net/

mation System,[2] the World Gazetteer,[3] the classifications of the United Nations Group of Experts on Geographical Names,[4] the Getty Thesaurus of Geographic Names,[5] and the ISO 3166-1 Country Codes.[6]

While simple gazetteer lookup has the advantage of being language-independent, more advanced algorithms consider lexical and structural linguistic clues, as well as contextual knowledge contained in the documents; e.g. dealing with ambiguity by removing stop-words, identifying references to people and organizations [22], and applying contextual rules like "single sense per document" and "co-occurring place names indicate nearby locations". For each identified reference, this process assigns a probability *P(name, place)* that a given name refers to a particular place [14]. The location that receives the highest probability is then assigned a canonical taxonomy node such as EUROPE/AUSTRIA/VIENNA; 48°14' N, 16°20' E.

2.1 Managing Geospatial Context

Ontologies play a key role in managing geospatial context. While conflicting definitions of "ontology" abound [23], most researchers agree that the term refers to a designed artifact formally representing shared conceptualizations within a specific domain [24, 25]. Deriving ontologies from unstructured textual resources [26] helps validate and extend domain-specific ontologies. Special emphasis should be placed upon integrating unstructured and structured information sources, and developing generic services encompassing both types of information. Ontologies support query term expansion and disambiguation, relevance ranking, and Web resource annotation. Visualizations of ontological structures can also serve as a navigational aid for the more experienced users.

Geo-ontologies encode geographical terms and their semantic relationships – e.g. containment, overlap, and adjacency [20]. In the case of spatially aware search engines, ontological knowledge supports query term expansion and disambiguation, relevance ranking, and Web resource annotation [27]. Geo-ontologies can either be expressed with the same markup language (e.g., OWL DL) as the domain ontology, or use the more specific Geographic Markup Language (GML) of the Open Geospatial Consortium [28].

2 Emerging Interface Technologies

Academia and industry alike call for advanced navigation and interaction facilities to access complex knowledge repositories, following the information seeking mantra of Shneiderman [29] and applying techniques from the emerging field of visual analytics [30]. A new generation of geospatial interface technologies addresses this call, facili-

[2] http://geonames.usgs.gov/
[3] http://www.world-gazetteer.com/
[4] http://unstats.un.org/unsd/geoinfo
[5] http://www.getty.edu/research/conducting_research/vocabularies/tgn/
[6] http://www.iso.org/iso/en/prods-services/iso3166ma/index.html

tating the access and manipulation of geospatial data. Online services such as *Google Maps*, *Yahoo! Local Maps*, and *MSN Virtual Earth (Windows Live Local)* are evolving quickly, adding new functionality, data sources and interface options in rapid succession. In the category of truly three-dimensional representations, dubbed the "holy grail of mapping" [31], the scale-independent spherical globes of Google Earth[7] and NASA World Wind [32][8] dominate the market. These geo-browsers aggregate and project layers of metadata onto cartographic displays.

Effectively navigating complex information spaces requires new interface metaphors. In conjunction with geospatially referenced information spaces, geo-browsers can be used to map annotated content units, track the virtual movements of individual users, and visualize the structure and dynamics of virtual communities. Diverting them from their traditional purpose and coupling them with semantically referenced information spaces, they can also be used to visualize and navigate "Knowledge Planets" based on layered thematic maps. Two distinct interface metaphors, the Earth and knowledge planets, allow users to rapidly switch between geospatial and semantic topologies, providing an unprecedented level of transparency and control over complex and heterogeneous datasets. The ease of switching between interface metaphors and the underlying context acknowledges the increasing importance of finding associations between information entities.

One-dimensional interfaces of text-based search engines and simple navigational systems are insufficient to visualize contextualized information spaces, as they would overwhelm users with an abundance of linear data. Besides, a reduction in dimensionality always goes hand in hand with a loss of information (e.g. projecting a 100-dimensional term space onto a two-dimensional visualization space) – the number of target dimensions positively correlates with the precision of the visualization.

Visualizing social interactions and the rapid diffusion of information increases the challenge of finding a suitable interface metaphor. As the concepts of "desktop", "village" and "landscape" have shown, well-known interface metaphors are powerful instruments to gain market acceptance [33]. Interface metaphors exploit the implicit processing capacity of the human's visual sense. They take advantage of the user's experience from previous perceptions [34], and her ability to navigate in two or three dimensions, recognize patterns, track movements, and compare objects of different size and color [35].

2.1 Geospatial Projections

NASA World Wind and Google Earth promote the "planet" metaphor by providing users with an accurate visual representation and allowing them to browse geospatial data from a satellite perspective. Most geo-browsers offer Application Programming Interfaces (APIs) or XML scripting to encourage developers of third-party online services [36]. Multiple layers of icons, paths and images can be projected via these services. Various visual elements are scaled, positioned on the globe, and linked to the resources of the contextualized information spaces such as Web documents, photo

[7] http://earth.google.com/

[8] http://worldwind.arc.nasa.gov/

collections, and personal contacts. A good example of using NASA World Wind to display additional layers of information is the Moderate Resolution Imaging Spectroradiometer (MODIS) service, providing daily updated planetary imagery at resolutions up to 250-meters-per-pixel, and documenting natural events such as fires, floods, storms, and volcanic activity.[9] The left screenshot of Figure 2 shows a MODIS overlay of Hurricane "Katrina" as of 29 August 2005.

The availability of APIs is largely responsible for the growing popularity of location-based services, often implemented as a mash-up that combines publicly available data and interface services from more than one provider into an integrated user experience [37]. The map in the center of Figure 2 displays the original Sigalert.com service that aggregates real-time traffic data from the San Francisco Bay Area. The screenshot on the right uses Sigalert.com data to visualize accidents and current traffic speeds onto the Google Earth representation of Southern Los Angeles.[10]

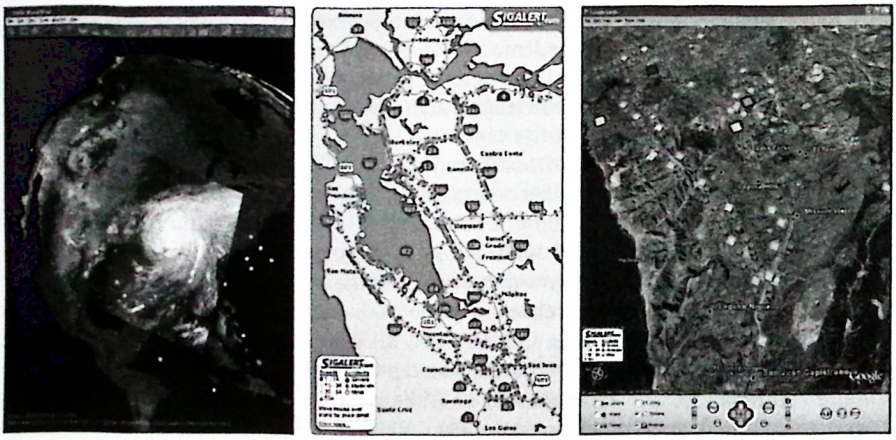


Fig. 1. Data Integration with NASA World Wind and Google Earth

2.2 Geo-temporal News Browsers

Hybrid models of individual and collaborative content production are particularly suited for geo-browsers, which can integrate and map individual sources (monographs, commentaries, blogs), edited sources (encyclopedias, conference proceedings, traditional newsrooms), evolutionary sources (Wiki applications, open-source project documentations), and automated sources (document summarizers, news aggregators). Geo-browsing technology not only impacts the production of content, but also its distribution, packaging, and consumption. When specifying preferences for personalized news services, for example, geo-browsers are effective tools to pinpoint locations and specify geographic areas to be covered.

[9] http://modis.gsfc.nasa.gov/

[10] http://bbs.keyhole.com/ubb/download.php?Number=75329

Personalized news services rely on content fragments annotated along semantic, spatial and temporal dimensions. The geographical scope of an article allows filtering and prioritizing electronic content in line with the user's current task, which can differ from her actual location. Topical similarity is another common dimension to tag and filter news content, often matched against user-specific degree of interest functions. Finally, by adding a temporal dimension (e.g. through time distribution graphs or visual animation), change over time along any other dimension can be captured: the unfolding of events, news distribution patterns, or the inter-individual propagation of messages. Dynamic queries, interactive time displays and playback controls will enable users to identify the rise and decay of topics – e.g. the diffusion of news coverage on natural disasters, or the popularity of specific tourism destinations.

The news summary on the left side of Figure 2 gathers the News Feeds of Associated Press,[11] processes them with the Yahoo! Geocoding API,[12] and displays the results via the Google Maps interface.[13] The second screenshot of Figure 2 shows a geo-temporal news browser that allows users to search a news database via query terms and time-interval sliders, and presents matching articles mapped onto a region of interest [38]. It follows Shneiderman's [29] information seeking mantra: provide an overview, allow zoom and filter, and present details on demand. These guidelines avoid clutter in the display, which results from projecting too many content items from a large knowledge repository simultaneously [19]. Instead of showing the complete set of available news items, for example, a user may wish to restrict the display to articles on climate change that were published in the online editions of Italian newspapers within the last 48 hours.

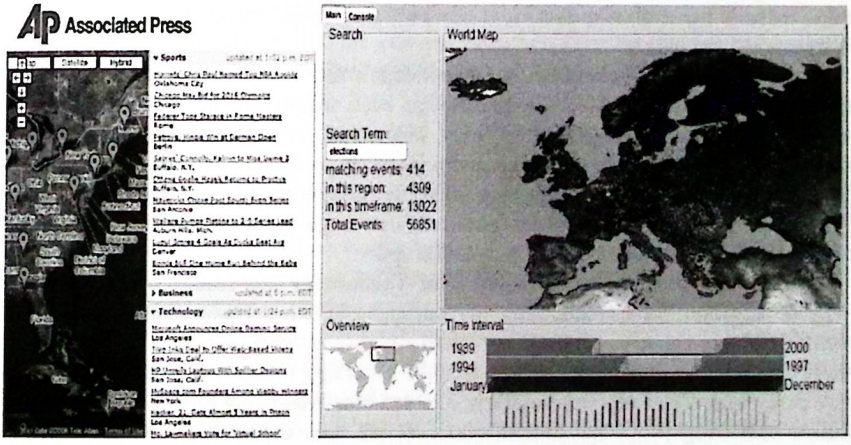


Fig. 2. Interfaces for Accessing Geo-Referenced News Archives

[11] http://hosted.ap.org/dynamic/fronts/RSS
[12] http://developer.yahoo.com/maps/rest/V1/geocode.html
[13] http://81nassau.com/apnews/

2.3 Knowledge Planets

Geo-browsers not only allow visualizing geographic data, but can also serve as a generic image rendering engine to project other types of imagery. Generally, two sets of information need to be integrated and mapped to latitude and longitude: image tiles and terrain information. Using information landscapes instead of cartographic data as the data source for the mapping process represents a radically new way of using geo-browsing technology. Implementations of thematic maps, i.e. visual representations of semantic information spaces based on a landscape metaphor [39], are visually compelling but often did not succeed beyond proof-of-concepts or early prototypes. The restriction to two dimensions and the significant overheads involved in developing and testing truly interactive client applications often kept researchers from disseminating their ideas to a wider audience.

Knowledge planets address both limitations. They can be generated by tiling thematic maps and orthographically projecting them onto the spherical globes of geo-browsers. When defining graphical excellence, Tufte recommends graphical displays that reveal the data at several levels of detail. He identifies layering and separation as the most powerful devices for reducing noise and enriching the content of displays [40]. The layered structure of knowledge planets, analogous to Landsat-7 data of different resolutions, follows this recommendation and allows visualizing massive amounts of textual data. The peaks of the virtual landscape represent focused media coverage on a particular topic, whereas valleys represent sparsely populated parts of the information space. Zooming provides an interactive, intuitive way of selecting the level of aggregation. Unique resource identifiers link concepts embedded in the thematic maps to external sources such as news articles, encyclopedia entries, and papers in scientific journals.

At the time of map generation, the knowledge planet's topology is determined by the structure and content of the knowledge base. Applying the planet metaphor, search results can be visualized as cities, landmarks, or other static objects of the man-made environment. With such a query interface, accessing and filtering complex data along multiple dimensions is as intuitive for users as using a geo-browser to get a glimpse of their next holiday destination. Knowledge planets hide the underlying complexity of the contextualized information space.

Previous research at Graz University of Technology has resulted in VisIslands [41], an algorithm for thematic mapping similar to SPIRE's Themescape [42] and its commercial successor Cartia/Aureka.[14] The algorithm, which supports dynamic thematic clustering of documents, can be adapted to serving image tiles for geo-browsing platforms. The knowledge repository is first pre-clustered using hierarchical agglomerative clustering [43]. The cluster centroids are then distributed randomly in the viewing rectangle. The documents belonging to each cluster, as determined by the initial pre-clustering, are then placed in a circle around each centroid. This arrangement is fine-tuned using a linear iteration force-directed placement algorithm adapted from Chalmers [44]. Shown in Figure 3, the result resembles a contour map of islands. Fortunately, algorithms based on force models easily generalize to knowledge planets' spherical geometries.

[14] http://www.cartia.com/static/aureka.htm

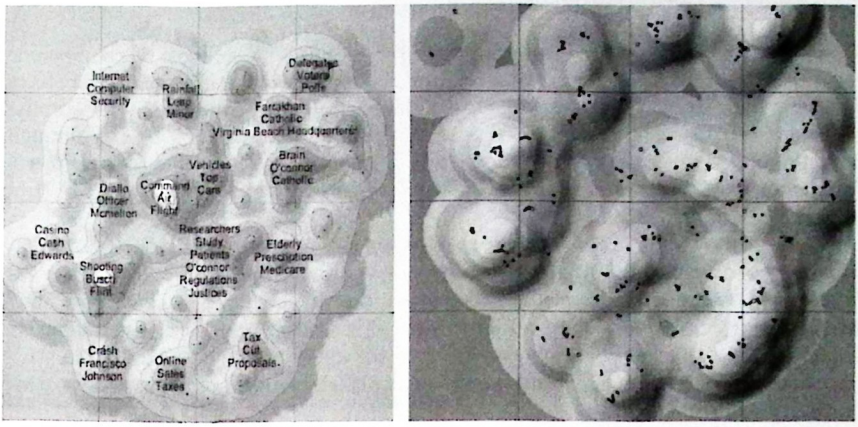


Fig. 3. Information Landscapes of Cartia/Aureka (left) and VisIslands (right)

Thematic mapping algorithms can be used to dynamically cluster and visualize search results while users navigate the contextualized information spaces, with new results being incorporated into the thematic landscape as they arrive. For very large document collections, the topology can be pre-computed and updated on a regular basis, for example overnight [45]. Based on query terms, search results are gathered from the knowledge repository, processed for keywords, clustered, and visualized. When users interactively explore the knowledge planet, they can display and compare additional properties of the search results (document source, date of publication, etc.). To refine their query, they simply have to choose the appropriately labeled clusters in the visualization, providing a quick and convenient way of specifying additional search requests.

The transition from two-dimensional thematic maps to knowledge planets poses a number of conceptual and technical challenges. The spherical globes of knowledge planets complicate the definition of the semantic topology – e.g. the initial arrangement of the major concepts (which will be guided by the domain ontology). Users will expect a consistent experience when rotating the planet. This requires a seamless flow of concepts when crossing the planet's 0° meridian line. The same principle applies to zooming operations. Multiple layers of thematic maps in different resolutions have to be synchronized with each other, and with the adaptive relevance ranking algorithms that avoid clutter by determining which and how many icons are incorporated into the display (clutter usually results from attempts to project too many content items simultaneously). Previous research has developed methods to automatically summarize, split or merging labels and other visual elements that tend to produce clutter [45].

On the server side, the number of layers as well as the possible resolution of the most detailed layer will largely be determined by the availability of computational resources. On the client side, geo-browsers gain from the wide-spread availability of high-end graphics cards for spherically projecting textures and vertices.

2 Tightly Coupled Interfaces

For online marketers, failing to respond to user queries with the most relevant results is a missed marketing opportunity. Traditionally, attempts to tackle this problem have focused on enhancing document classification and ranking algorithms. More recently, the field has started to look at methods to help users improve their queries through disambiguation and query refinement. Domain-specific search engines [46] are therefore becoming increasingly popular because they enable to interpret and refine queries with increased accuracy.

Figure 4 shows the sketch of a browser extension, which aims to reduce the cognitive overhead caused by a lack of semantic, spatial and social context [47]. The *detailed view* on the left side displays the currently active document – either in reading mode (e.g. accessing a Web page), or in writing mode (e.g. collaboratively authoring a document). Context-providing *overviews* in the upper right side visualize the position of this document within the knowledge planet's semantic space, and its geographic positioning.

Below the overviews, just-in-time information retrieval agents generate a summary of related documents, as well as a list of other users associated with the document's content. Users of just-in-time information retrieval agents have been found to be more efficient at retrieving information, and to incorporate more information into their daily workflows that they would with traditional search engines [48]. To aid users in remembering their location, a "you are here" indicator is embedded in the overviews in the upper right corner, as well as an ontology-based navigational aid (the small hierarchical diagram on the right side of the screen).

The active document needs to be tagged automatically and in real time. If a user edits the active document, for example, the position of the "you are here" indicators in the overviews should reflect these changes immediately. Geo-tagging allows the browser extension to project the document onto the correct physical location. Semantic tagging allows the browser extension to (i) rotate the knowledge planet accordingly and mark the document's position on its surface; (ii) update the ontology-based navigational aid. Browsing the Web (i.e. changing the detailed view) thus updates the position of the indicator in the overview. Hence, the various views are said to be tightly coupled [49, 50].

Such *tightly coupled views* are straightforward to implement and understand. User actions in either part of the window cause modifications in all the other related display. The continuous and synchronized display of several views on the contextualized information space will allow rapid and reversible interaction. The advantages and positive effects of tightly coupled interface components on user performance have been documented in several studies [51-53]. Tightly coupled interfaces reduce search time, allow the detection of patterns, and aid users in choosing the next node to visit [54, 3, 55].

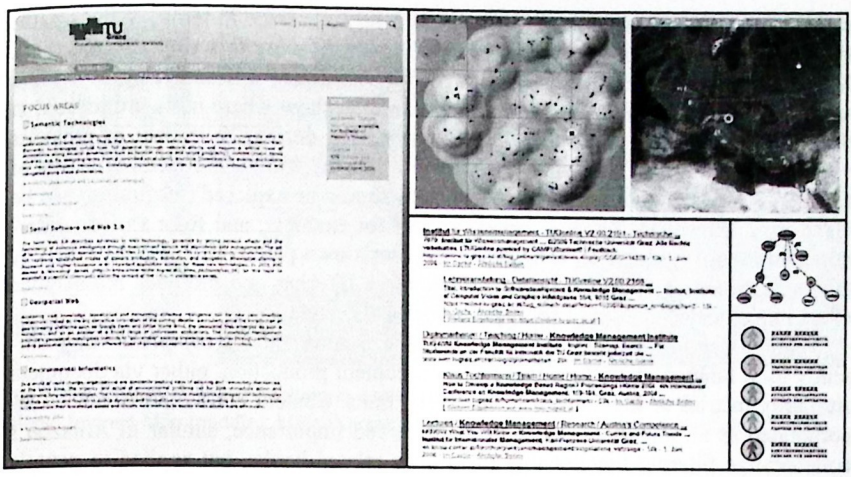


Fig. 4. Sketch of a tightly coupled view showing a rendered document, two distinct overviews (knowledge planet and geospatial projection), a small ontology-based navigational aid, and two views for the just-in-time retrieval of related Web resources and user profiles

The screenshot of Figure 4 is just one possibility of composing various interface services into an integrated user experience on a single 16:9 screen. Modular service architectures allow adapting the layout depending on the requirements of a specific application, or individual preferences. Users not interested in the geospatial context of their activities, for example, can choose to replace the geographic overview with a more detailed rendering of the ontology. For the most demanding users, a multiple screen setup should be most appropriate: the usual desktop environment on the center screen (e.g. document, ontology-based navigational aid, and just-in-time information retrieval agents for related documents, RSS feeds and blogs), and two separate screens for depicting Earth and the knowledge planet, respectively.

3 Visualizing User Activity

Observing, aggregating and visualizing human behavior is a common activity [50], in the past often restricted by the availability of technology. The left diagram of Figure 5 exemplifies the labor-intensive customer tracking typical for traditional retailing outlets in the 1970s (Becker, 1973). Electronic collaborative environments do not need to track the movements of users in a separate process, since their movements and interactions are an integral part of the contextualized information space. These interactions can be regarded an implicit source of user feedback. Explicit feedback is gathered during the interaction in the form of user reviews or recommendations (content, locations, expertise of other users, etc.), and during the formal user interface evaluation, which is independent from the system's actual operation.

As outlined in the preceding section, the knowledge planet's topology is determined at the time of map generation. Similar to Certau's *Wandersmänner [56]*, users represent a collective and virtual presence hovering above this topology. Their

movements can be traced on the knowledge planet's surface in terms of their paths (here well-trodden, there very faint) and trajectories (going this way and not that). Knowledge planets are excellent platforms for visualizing and analyzing individual and aggregated user interactions. Not only can they show where users are located at the moment, but also where they have already been during the current or previous visits (their "footprints" or "trails"). Extensions like the specific form and direction of footsteps, or dust clouds raised by recent steps should be explored (inspiration can be taken from the study of animal behavior, ants for example, and from various entertainment industry products). Comparing the user's own path with the paths taken by others represents a special type of collaborative filtering and provides reassurance when implemented as a supplemental navigational system.

Analytic services should consider such social information, complementing currently used ranking algorithms that focus on content production, either via document similarity metrics or the number of incoming links. Content consumption can be incorporated as an important indicator of perceived importance, similar to Amazon's collaborative filtering functionality to identify related books, but applied to generic content items. The incorporation of social behavior also resonates well with the Web 2.0 philosophy, and the idea of harnessing collective intelligence.

The same principle not only applies to the individual users of the system, but also to the analysts interested in macro-level effects like the overall patterns in search behavior, or the most popular concepts in a given timeframe. Aggregated representations such as Telegeography's Global Internet Map[15] shown in the middle of Figure 5, however, are often static in nature. User behavior tends to fluctuate heavily, and thus requires more dynamic visual metaphors. Weather maps seem an ideal candidate, as they are a proven method to condense highly complex and dynamic datasets into a visual representation that is useful for experts and understandable for general public at the same time. Techniques developed for weather maps could be used for both local and global trends in the users' interactions. Clouds might represent their aggregated search behavior, for example, while isobars highlight areas of similar interest. Animated isobars would be ideal to visualize trends in content consumption. Those "iso-interest contours" would connect resources sharing the same level of user interest. Clear and sunny high pressure areas would contain topics of limited interest to the average user, while turbulent low pressure systems would indicate heightened levels of activity. Thunderstorms would symbolize the current hotspots of user activity. Many of these hotspots would be triggered by intensive media coverage on exogenous events, and subsequent discussions; e.g. natural disasters such as last year's hurricane "Katrina", and the recent earthquake which struck central Java. Other hotspots might emerge through a process of resonance, when insignificant events or social interactions trigger massive reactions within the contextualized information space.

This opens up interesting analytical opportunities, when potential causes of observed user behavior are mapped against that behavior over time. Isobars could represent real-world data, and color-coding the users' behavior. Then the isobars most closely resembling the color-coding would be most likely to be responsible for the trend in question.

---

[15] http://www.telegeography.com/

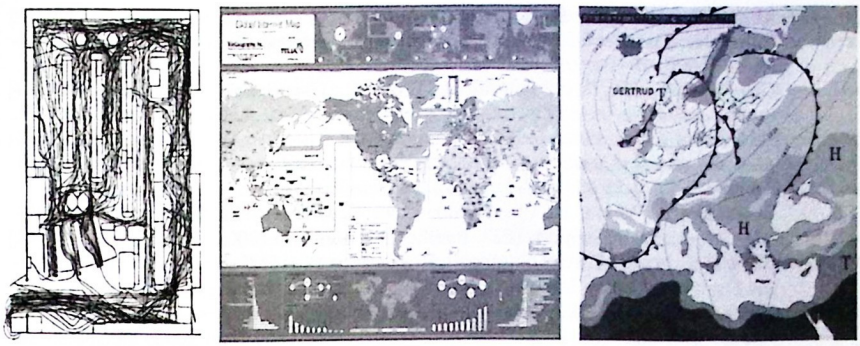


Fig. 5. Traditional customer tracking in a retailing outlet (left); Telegeography[16] Global Internet Map (middle); MeteoXpress[17] weather map as of 22 May 2006 (right)

5 Summary and Outlook

The geo-tagging of existing resources is rapidly gaining the attention of researchers and practitioners alike, because geography emerges as a fundamental principle for structuring complex knowledge repositories, yielding the world's knowledge through the lens of location [31, 58]. Adding location metadata to existing resources and using geo-browsers to access these resources weds physical and virtual spaces. This integration deepens our experiences of information spaces and incorporates them into our every-day lives [36].

This paper presented a conceptualization of advanced collaborative technology based on content, annotation and ontology services to build contextualized information spaces, interface technologies that let users choose between semantic and geospatial topologies for intuitively navigating these spaces, and a framework for analyzing content diffusion and interaction patterns within interactive environments.

References

[1] Erle, S., Gibson, R., Walsh, J.: Mapping Hacks - Tips & Tools for Electronic Cartography. O'Reilly, Sebastopol (2005)

[2] Udell, J.: Annotating the Planet with Google Maps. InfoWorld March 04, 2005 (2005) http://www.infoworld.com/article/05/03/04/10OPstrategic_1.html

[3] Benjamins, R., Contreras, J., Corcho, O., Gómez-Pérez, A.: Six Challenges for the Semantic Web. AIS SIGSEMIS Bulletin 1 (2004) 24-25

[4] McCurley, K. S.: Geospatial Mapping and Navigation of the Web. In: Proc. 10th International World Wide Web Conference (2001) 221-229

[16] http://www.telegeography.com/

[17] http://en.meteoxpress.com/

[5] Motta, E., Shum, S. B., Domingue, J.: Ontology-Driven Document Enrichment: Principles, Tools and Applications. International Journal of Human-Computer Studies 52 (2000) 1071-1109

[6] Ciravegna, F., Dingli, A., Petrelli, D., Wilks, Y.: User-System Cooperation in Document Annotation Based On Information Extraction. In: A. Gomez-Perez and V. R. Benjamins, (eds.): Knowledge Engineering and Knowledge Management - Ontologies and the Semantic Web (Lecture Notes in Artificial Intelligence, Vol 2473). Springer (2002) 65-80

[7] Domingue, J., Motta, E.: PlanetOnto: From News Publishing to Integrated Knowledge Management Support. IEEE Intelligent Systems 15 (2000) 26-32

[8] Daviel, A., Kaegi, F. A.: Geographic Registration of HTML Documents (IETF Internet-Draft, July 2003). Internet Engineering Task Force, Sterling (2003)

[9] Jones, C. B., Alani, H., Tudhope, D.: Geographical Information Retrieval with Ontologies of Place. In: D. R. Montello, (ed.) International Conference on Spatial Information Theory: Foundations of Geographic Information Science (= Lecture Notes in Computer Science, vol. 2205). Springer (2001) 322-335

[10] Ding, J., Gravano, L., Shivakumar, N.: Computing Geographical Scopes of Web Resources. In: Proc. 26th International Conference on Very Large Data Bases (2000) 545-556

[11] Delboni, T. M., Borges, K. A. V., Laender, A. H. F.: Geographic Web Search based on Positioning Expressions. In: Proc. 2nd International Workshop on Geographic Information Retrieval (2005) 61-64

[12] Morimoto, Y., Aono, M., Houle, M. E., McCurley, K. S.: Extracting Spatial Knowledge from the Web. In: Proc. Symposium on Applications and the Internet (2003) 326-333

[13] Bell, B.: Vienna Marking Mozart Milestone. (2006) http://news.bbc.co.uk/2/hi/entertainment/4654880.stm

[14] Amitay, E., Har'El, N., Sivan, R., Soffer, A.: Web-a-Where: Geotagging Web Content. In: Proc. 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (2004) 273-280

[15] Wang, C., Xie, X., Wang, L., Lu, Y., Ma, W.-Y.: Detecting Geographic Locations from Web Resources. In: Proc. 2nd International Workshop on Geographic Information Retrieval (2005) 17-24

[16] Cowie, J., Lehnert, W.: Information Extraction. Communications of the ACM 39 (1996) 80-91

[17] Weiss, S. M., Indurkhya, N., Zhang, T., Damerau, F. J.: Text Mining - Predictive Methods for Analyzing Unstructured Information. Springer, New York (2005)

[18] Kienreich, W., Granitzer, M., Lux, M.: Geospatial Anchoring of Encyclopedia Articles. In: Proc. 10th International Conference on Information Visualisation (2006) In Print

[19] Larson, R. R.: Geographic Information Retrieval and Spatial Browsing. In: L. Smith and M. Gluck, (eds.): GIS and Libraries: Patrons, Maps and Spatial Information. University of Illinois (1996) 81-124

[20] Tochtermann, K., Riekert, W.-F., Wiest, G., Seggelke, J., Mohaupt-Jahr, B.: Using Semantic, Geographical, and Temporal Relationships to Enhance Search and Retrieval in Digital Catalogs. In: Proc. 1st European Conference on Research and Advanced Technology for Digital Libraries (= Lecture Notes in Computer Science, Vol 1324) (1997) 73-86

[21] Hill, L. L., Frew, J., Zheng, Q.: Geographic Names - The Implementation of a Gazetteer in a Georeferenced Digital Library. D-Lib Magazine 5 (1999) www.dlib.org

[22] Clough, P.: Extracting Metadata for Spatially-Aware Information Retrieval on the Internet. In: Proc. 2nd International Workshop on Geographic Information Retrieval (2005) 25-30

[23] Guarino, N.: Understanding, Building and Using Ontologies. International Journal of Human-Computer Studies 46 (1997) 293-310

[24] Jarrar, M., Meersman, R.: Formal Ontology Engineering in the DOGMA Approach. In: R. Meersman and Z. Tari, (eds.): International Conference on Ontologies, Databases and Applications of Semantics (Lecture Notes in Computer Science, Vol 2519). Springer (2002) 1238-1254

[25] Gahleitner, E., Behrendt, W., Palkoska, J., Weippl, E.: Knowledge Sharing and Reuse: On Cooperatively Creating Dynamic Ontologies. In: Proc. 16th ACM Conference on Hypertext and Hypermedia (2005)

[26] Liu, W., Weichselbraun, A., Scharl, A., Chang, E.: Semi-Automatic Ontology Extension Using Spreading Activation. Journal of Universal Knowledge Management 0 (2005) 50-58

[27] Abdelmoty, A. I., Smart, P. D., Jones, C. B., Fu, G., Finch, D.: A Critical Evaluation of Ontology Languages for Geographic Information Retrieval on the Internet. Journal of Visual Languages and Computing 16 (2005) 331-358

[28] Lake, R., Burggraf, D., Trninic, M., Rae, L.: Geography Mark-Up Language: Foundation for the Geo-Web. John Wiley & Sons, New York (2004)

[29] Shneiderman, B.: Designing the User Interface: Strategies for Effective Human-Computer Interaction. Addison-Wesley, Berkeley (1997)

[30] Wong, P. C., Thomas, J.: Guest Editors' Introduction: Visual Analytics. IEEE Computer Graphics and Applications 24 (2004) 20-21

[31] Levy, S.: Making the Ultimate Map. Newsweek 143 (2004) 56-58

[32] Hogan, P., Kim, R.: NASA Planetary Visualization Tool. In: Proc. American Geophysical Union Fall Meeting (2004)

[33] Fidler, R.: Mediamorphosis: Understanding New Media. Pine Forge Press, Thousand Oaks (1997)

[34] Purves, D., Lotto, R. B., Williams, S. M., Nundy, S., Yang, Z.: Why We See Things the Way We Do: Evidence for a Wholly Empirical Strategy of Vision. Philosophical Transaction of the Royal Society of London 356 (2001) 285-297

[35] Le Grand, B.: Topic Map Visualization. In: J. Park and S. Hunting, (eds.): XML Topic Maps - Creating and Using Topic Maps for the Web. Addison-Wesley (2003) 267-282

[36] Roush, W.: Killer Maps. Technology Review 108 (2005) 54-60

[37] Hof, R. D.: Mix, Match, And Mutate: "Mash-ups" – Homespun Combinations of Mainstream Services – are Altering the Net Business Week 3942 (2005) 72

[38] Rüger, S.: Putting the User in the Loop: Visual Resource Discovery. In: Proc. 3rd International Workshop on Adaptive Multimedia Retrieval (2005)

[39] Chalmers, M.: Using a Landscape Metaphor to Represent a Corpus of Documents. In: A. U. Frank and I. Campari, (eds.): Spatial Information Theory: A Theoretical Basis for GIS (Lecture Notes in Computer Science, Vol 716). Springer (1993) 377-390

[40] Tufte, E. R.: Envisioning Information. Graphics Press, Cheshire (1990)

[41] Andrews, K., Guetl, C., Moser, J., Sabol, V., Lackner, W.: Search Result Visualisation with xFIND. In: Proc. Second International Workshop on User Interfaces to Data Intensive Systems (2001) 50-58

[42] Wise, J. A.: The Ecological Approach to Text Visualization. Journal of the American Society for Science 50 (1999) 814-835

[43] Jain, A., K., Murty, M. N., Flynn, P. J.: Data Clustering: A Review. ACM Computing Surveys 31 (1999) 264-323

[44] Chalmers, M.: A Linear Iteration Time Layout Algorithm for Visualising High-Dimensional Data. In: Proc. 7th Conference on Visualization (1996) 127-132

[45] Andrews, K., Kienreich, W., Sabol, V., Becker, J., Droschl, G., Kappe, F., Granitzer, M., Auer, P., Tochtermann, K.: The InfoSky Visual Explorer: Exploiting Hierarchical Structure and Document Similarities. Information Visualization 1 (2002) 166-181

[46] McCallum, A., Nigam, K., Rennie, J., Seymore, K.: A Machine Learning Approach to Building Domain-Specific Search Engines. In: Proc. 16th International Joint Conference on Artificial Intelligence (1999) 662-667

[47] Turoff, M., Hiltz, S. R., Bieber, M., Fjermestad, J., Rana, A.: Collaborative Discourse Structures in Computer Mediated Group Communications. In: Proc. 32nd Hawaii International Conference on System Sciences (1999)

[48] Rhodes, B. J., Maes, P.: Just-in-Time Information Retrieval Agents. IBM Systems Journal 39 (2000) 685-702

[49] Kumar, H. P., Plaisant, C., Shneiderman, B.: Browsing Hierarchical Data with Multi-level Dynamic Queries and Pruning. International Journal of Human-Computer Studies 46 (1997) 103-124

[50] Scharl, A.: Explanation and Exploration: Visualizing the Topology of Web Information Systems. International Journal of Human-Computer Studies 55 (2001) 239-258

[51] Chimera, R., Shneiderman, B.: An Exploratory Evaluation of Three Interfaces for Browsing Large Hierarchical Tables of Contents. ACM Transactions on Information Systems 12 (1994) 383-406

[52] Norman, K., Weldon, L., Shneiderman, B.: Cognitive Layouts of Windows and Multiple Screens for User Interfaces. International Journal of Man-Machine Studies 25 (1986) 229-248

[53] Shneiderman, B., Shafer, P., Simon, R., Weldon, L.: Display Strategies for Program Browsing: Concepts and an Experiment. IEEE Software 3 (1986) 7-15

[54] Nielsen, J.: Multimedia and Hypertext: The Internet and Beyond. AP Professional, Boston (1995)

[55] Card, S. K., Mackinlay, J. D., Shneiderman, B.: Readings in Information Visualization: Using Vision to Think. Morgan Kaufmann, San Francisco (1999)

[56] De Certeau, M.: The Practice of Everyday Life. University of California Press, Berkeley (1983)

Acknowledgement

IDIOM (Information Diffusion across Interactive Online Media, www.idiom.at) is funded by the Federal Ministry of Transport, Innovation and Technology (BMVIT) and the Austrian Research Promotion Agency (FFG) within the strategic objective FIT-IT (www.fit-it.at). The author wishes to thank his colleagues at the Know-Center and at Vienna University of Economics and Business Administration for their valuable feedback on earlier versions of this manuscript. The Know-Center is funded by the Austrian Competence Center program Kplus under the auspices of the Austrian Ministry of Transport, Innovation and Technology (www.ffg.at), and by the State of Styria.

Matching-concepts: An ontology Approach to Retrieve Geographical Information

Felix Mata[1] and Serguei Levachkine[2]

Geoprocessing Laboratory (GEOLAB)
Centre for Computing Research (CIC), National Polytechnic Institute (IPN)
Av. Juan de Dios Bátiz s/n Unidad Profesional "Adolfo López Mateos"
07738, México, D.F., MEXICO
migfel@sagitario.cic.ipn.mx[1], sergei@cic.ipn.mx[2]

Abstract. Nowadays one of the research directions in Geographical Information Retrieval (GIR) is focused on finding information related to particular geographical locations. Some approaches have treated this problem using geo-referenced information available from several sources, e.g. digital maps and spatial databases. However, reported in literature results are not very relevant to the user's expectations, because the free-text queries have computed without considering properties and relations between geographic objects, e.g. topology of spatial data. If users express queries by names of places, prepositions, e.g. "near", "along", etc., then we will require additional mechanisms, e.g. spatial semantics treatment, to catch the nature of geographic objects appeared in these queries. In this work, we propose a method to match concepts using ontologies. In other words, our approach to the information retrieval is based on the spatial semantic properties and relations between objects and not only on text analysis. This approach improves the relevance of results for free-text queries, which contain geographic or spatial objects. We describe a method to retrieve spatial information based on the semantics of geographic components presented in a free-text query. Our approach represents an alternative to the keyword-matching. The idea consists of extracting the concepts presented in queries, and then matching them to the corresponding contexts. These contexts are embedded in ontologies. The best matching concept is obtained when the related to it contexts converge within an ontology. We semiautomatically built ontology from a universe of documents and propose a method to scan the ontology until finding the suitable concept-matching in accordance to the context of free-text query on a specific domain. In particular, we use queries of location and position. We present some results and show that the relevance of these results is closer to the user's expectations. Finally, we compare our approach to the approaches that are not considered the semantics of spatial data.

1 Introduction

Presently, the retrieval of geographic information related to particular geographical locations is based on geo-referenced information sources (i.e., linked information to geographic coordinates) available on several sources (maps, geographically indexed

© M. Moreno-Ibarra, S. Levachkine (Eds.)
Semantics in Geoinformatics Applications
Research in Computing Science 25, 2006, pp. 19-33

books, etc.) But, people do not make queries using geographic coordinates, very often a search starts submitting free-text queries to database systems, search engines or Web information systems. A free-text query contains several terms related to geographical or spatial objects, but not in the way that a GIS defines operations with spatial data, for example: within or inside can be interpreted as synonyms, but these can represent a buffer operation or other topological operation, the decision of which chosen can be solved by semantics on the query. With this sample we can see the importance of considering the semantics in free-text queries that involve geographic objects.

The goal of a search is to find information about any subject or activity related with our daily life, although commonly the results obtained do not satisfy the expectations of users. This problem occurs mainly because the techniques to solve queries are based on exact-match or keyword-match, these techniques are used in database systems and search engines respectively. So, these approaches present some problems such as the ambiguity of words.

Here is where others research lines have built techniques to avoid or solve these problems, one of them is the information retrieval (IR) where the answer to queries are based on the approximations to expected results. These expected results are denominated as *relevant,* additionally; the results are not ordered but ranked, by means of several methods (most of them using syntax techniques).

The main goal of IR consists of retrieving text documents that are relevant to a given query, where a document is considered relevant when it contains one or several words that also appear in queries. Nevertheless, that approach does not ensure finding the suitable answer, because the lack of search terms in some documents does not necessarily mean that the documents are not relevant. Besides, another disadvantage of current IR systems is that they are based on exploiting the nature of text, therefore the semantics of a query cannot be considered. Therefore alternative methods to improve that retrieval process are required, for example considering context, semantics, etc. The case of spatial data is not the exception to the problem described below, besides the methods used in traditional information retrieval systems (IRS) are not appropriate to geographic information, because if we consider that the classical information retrieval models (e.g., Vector Space, Probabilistic, Boolean) [14] are based on lexicographic term matching, then there is no way to consider relationships or properties of geographic objects (required to extract the semantic of geographic query).

Although, there are approaches in which the searches are performed with the aid of ontologies [1], the used approach cannot be applied in Geographical information systems, because the nature of spatial data requires a special treatment [2]. For example, two terms can be semantically different although they are lexicographically similar (near is semantically different when the topography is considered and when it is not considered). Then, retrieving documents by classical retrieval methods will fail, if the semantics of terms is not considered. Besides, in GIS the semantic processing approach has been widely used and proves significant results [18, 19] in that way semantic processing promises to be an interesting alternative to Retrieval information focused in geographical aspects, this field of research is known as Geographic Information Retrieval.

Therefore, at this point our work is addressing: to retrieve information by extracting the semantics implicit on free-text queries, and matching concepts on ontology, where we test our approach using location queries (e.g "the bank near the conference", "hotel 1km of walking distance")

1.1 BACKGROUND AND RELATED WORK

Nowadays, the search process is a fundamental component of retrieval information systems, the case of geographic information retrieval (GIR) is not the exception where several works have been oriented to searches, in this direction one of the main trends consists of improving the techniques and algorithms to extract relevant information (*best* answers to certain queries) where we find enough efforts to achieve that goal, for example in [3] the GIR is performed by means of spatial Bayesian algorithms, focused on workspaces of a commercial GIS, where a workspace is integrated by several layers of spatial data, the idea is to extract the relations present between these layers , in order to be exploited by the retrieval process .

Others' works are addressed to the web, for example in [4] a geographic search was proposed using query-expansion by means of an API provided by a keyword-web search engine, one of the disadvantages of this approach is that expansion of queries (number of terms) is constrained by a search engine, then the retrieval process is affected in a negative way when many terms are required. The expansion approach has offered good results only in particular sceneries, where the number of terms was small, but when the number of terms is bigger, additional inconvenient and problems are presented, then the solution become another problem.

Other proposals are focused on solving the problem of ambiguity of words; the proposed solutions are based on a knowledge representation, such as: hierarchies of terms, taxonomies, and ontologies, but most of them are solutions based on text or syntax properties, while others describe treatment semantics without considering spatial relations, as in [5].

Into of this group several semantic approaches have been proposed, where one of the main contributions consists of including ontologies and semantic annotation, an example is described in [6].

The ontologies [7, 14] have been widely used in several semantic approaches; they are applied in, practically, any domain and of course in GIS field [8, 15]. Nevertheless, these approaches not consider processes and algorithms to explore ontologies, because an ontology describes domain theories for the explicit representation of the semantics of data [9], then we can use the ontology to know the semantics of query. Then, we need algorithms to explore these ontologies and getting the semantics required. In [10] the authors propose a way to match a location expression with certain places. It process is done using a database of places, the database is divided in two groups, the first one is formed by terms driven popularly and the second one is formed by terms very specialized. They show how geocoding can be implemented over incomplete and possibility inaccurate addressing data. Additionally, the paper shows a way to treat qualitative and quantitative data, but as a disadvantage a geocoding process is required previously. Other approaches are focused in ranking algorithms where very often the parameters considered are:

number of visits to sites and pages, or links and references, etc. An example is [11] where geo-data are used. Then, it is the point where our approach is addressed, no use geocoding, because not all users can express in explicitly way the geographical coordinates for a site or place (location queries). Moreover, a significant growth of geographically oriented search, local search, is undeniable as is described in [17], and then we need to develop better mechanisms to solve the location queries without expressing the places' coordinates. Furthermore, Many works have dealt on unstructured information where data contains some geographic reference for example in [23, 24].

Additionally, although the GIS research community have suggested and made emphasis on the use and treatment of spatial relations, few studies have been addressed in that direction, some examples are published on [12,13] while that recent work has focused on qualitative spatial reasoning, a sample can be found in the often-cited model of topological relations among point sets [18].

In the field of retrieval information many models have been used, where the term-based Vector Space Model (VSM) is the state-of-the-art document retrieval method [16] and it is based on lexicographic term matching. While that in our approach the matching is performance by concept. In particular Information Extraction (IE) and Information Retrieval (IR) are used in conjunction to built new tools that offer better results in the search process [13]. The first one has the task of organize indexes to be exploited by IR. And the second one, IE can be described as the process of populate a structured information repository (index) from an un-structured information source [21]. That task is performance in automatic way in two modalities: the first one extracting either whole information of a document, where every term of a document is treated and a weight is associated to each term [7]. While that in second one some fragments of document are extracted using predefined rules to find out specific information [21]. IR works with models, techniques, mechanisms to extract information that has already been processed, stored and (e.g. plain text files, databases, XML files). In IR the fast processing of queries is possible because the index structure was previously built [22].

Our work is addressed in *geographic queries* in particular those that include terms of proximity, i.e. *near, distance, behind, at the side of , in,* where these terms can be interpreted based on their meaning and in that form to offer a relevant answer. Moreover, we consider frame reference (those that helps to denote a location or position), i.e. *"within easy walking distance to Azteca Stadium", "five minutes from Art museum".*

The remainder of this paper is structured as follows: Section 2 describes the process to extract the semantics of queries. Section 3 shows the semi-automatic process to built ontology. In section 4 an overview of retrieval and search strategy (matching concepts) is described, where some examples are presented. In Section 5 preliminary results are presented. Finally, some conclusions are drawn and we discuss possibilities for future work.

2 SEMANTICS EXTRACTION OF QUERIES

Very often people talk and write about places using references or approximations to describe locations, ways, routes, etc. But, these descriptions lacks of precision, the following expression is a clear example of that: "*The restaurant is **in front of** Mayan Hotel*"in that case, if the direction is not expressed, the meaning of ***in front of*** can be different depending on the user's position. Other sample occurs when place-approximations are used such as: "*The school is **near** main Avenue*" the meaning of ***near*** can change if the topography is considered or if is not considered. In these expressions a common characteristic is the use of a spatial relation (e.g *in front of*) and reference frames, or places widely known.

They are used for offering a major detail or precision about the location. These places widely known and spatial relations are rich on semantics, they can be extracted to help in the retrieval process, but a previous knowledge is required.

This previous knowledge is about the spatial relations and the places widely known (that knowledge can be got from ontology, dictionaries and gazetteers)

We explain that with the following scenario: suppose that a speaker wants to know if there are hotels near the conference venue, if he use a traditional system (e.g. Google) the results will be only documents that contain one or more words of the location query, but if we consider the possible relations associated to "*near*", we will have, additionally, the relations and properties needed to contextualize the search.

These relations and properties are extracted from dictionaries, gazetteers and ontology. In the case of ontology, they were extracted by exploration of it, while that in the case of dictionaries and gazetteers, will be extracted by sentence analysis of definitions by each term.

The ontology exploration consists of finding (matching) the query's terms in the ontology nodes. If the terms are found in the ontology, the relations and properties from them are extracted.

In the table 1 we can find the properties and relations extracted, for each term, from dictionaries and gazetteers to the location query: *"hotels **near** conference venue"*. In section 3.1 the extraction process is described. While that in section 2 the ontology construction process is explained.

Table 1. Relations and properties extracted to query :*"hotels near conference venue"*.

QUERY: *"HOTELS NEAR CONFERENCE VENUE".*	
CONCEPT	PROPERTIES AND RELATIONS
NEAR	Within **reach**
	Related to center
	Related to **periphery**
	Related to **time**
	Related to **place**
	Relies in some sort of **distance**
	Synonyms: close, approximately, vicinity
HOTEL	Related to **Tourism**
	Related to services

	Is a Place
	Related to **lodging**
	Has rooms
	Is a Resort
	Synonyms: guesthouse, boarding house
CONFERENCE VENUE	Related to **meeting**
	Related to **Topic**
	Related to **Event**
	Has schedule
	Has location
	Is a Place
	Synonyms: place of trial, setting of event

The relations and properties shown in the table 1 were used to construct the ontology, the bold font indicates that they are *concepts* (have relations and properties). Previously ,the *concepts* were manually processed, therefore, the ontology contains them (by definition ontologies are built based on concepts and relations) where that ontology will be enriched with additional relations (obtained from documents) during the rest of process. Moreover, the ontology will be used in conjunction with traditional retrieval approach to offer better results compared with those obtained by traditional systems.

3 THE ONTOLOGY CONSTRUCTION

In this section, we will explain the process to build the ontology. Basically, it is an approach based on two methods to achieve the goal. The first one is addressed to document collection and the second one is addressed to location query.

We processed the query and the documents employing these methods. In the case of documents we use an extraction method based on surrounding terms (EMST) and to the case of the query we apply an extraction method based on concept (EMC). The process was performance in semi-automatic way, using a program to extract the terms based on the two methods, afterward manually (according to the experience of specialist GIS) some terms are selected to be included in the ontology. The methods CEM and EMST are described in sections 3.1 and 3.2 respectively.

3.1 CONCEPTS EXTRACTION METHOD (CEM)

The method consists of two steps: the first one, where the terms are extracted from query (verbs and nouns) and stored in a table. The second one: dictionaries and gazetteers are used to find-extract the relations and properties of each concept. For example, for the earlier query we have the term *hotel* where the associated definition (from dictionary) is: *"is a place which provides overnight lodging and offers other services"*. In that case, the verbs are identified as relations, while that the nouns allow

identifying properties related to the concept. Then, a preliminary output for that process is:

Relations = {provide, is a, offer} Properties related= {place, services, overnight, lodging}

Here, the sentence analysis is based on mechanisms used in traditional IR, where the stopwords are not considered, but additionally rules are added to identify relations such as:*" is a" " has a" " part of*" etc.

3.2 EXTRACTION METHOD BY SURROUNDED TERMS (EMST)

This method is based on the hypothesis of that the surrounded terms to location query's terms are useful to find other relations. For example: a surrounded term to ***near*** could be subway station, street, or suburb, in each case we have a different representation (Point, Line, or Polygon respectively) therefore the relations will be different for each representation.

Another example is when a user submits the query: *"rivers in Mexico"*, then suppose that we have a document that contains the following fragment of text: *"Colorado **River** is a **river** in the south western United States and north western Mexico"*.

The surrounded terms (ST) to the "*River*" term are: ST= {in, is a, Colorado}
These ST will be added to the table that contains the results obtained using the CEM method. We explain now the overall process of EMST: First, based on document collection from previous experiment where the documents contain at least one proximity term or names of places widely known in Mexico City.

We computed the higher term frequency (HTF) and inverted document frequency (IDF) for each concept contained in each docut. The HTF and IDF is computed only to spatial relations (related with proximity) and the surrounded words to them (verbs and nouns) for each document. The *stopwords* were not considered. The Table 2 shows the properties and relations extracted to concept "*near*" using EMST and EMC.

Table 2. Semantics extracted to "near"

CONCEPT	SET OF SURROUNDED WORDS (SSW) AND HTF	PROPERTIES AND RELATIONS
NEAR	Along a route, Located, located inside, nearest, quite close to, conveniently located, located right, relatively close, Surrounded, 6 miles from, ~15 minutes from, is less than 10 minutes from, is situated, Only 1 km, less 10 minutes away from,…	Within **reach**
		Related to center
		Related to **periphery**
		Relies in some sort of **distance**
		Synonyms

Then, using these results the ontology is enriched by adding these relations and properties. The figure 1 shows a fragment of the ontology.

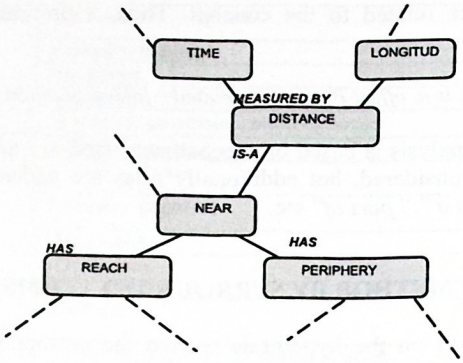


Fig. 1. Snippet of Ontology

As the Fig. 1 shows, there are properties and relations that can aid in the process of retrieval to location queries related to concept *near*, and not only by syntax properties. For example, in the table 2 the term distance appears in bold because *distance* is a concept in our ontology, therefore the documents that contain the concept *distance* will be processed, and also the *near* concept (because they have a relationship in the ontology).

Then, in this form, the relevant documents extraction for concept *near* will be performance by means of *distance* concept (without ontology, it would not be possible). In the following section we show an overview of the retrieval process, and a brief description of the strategy of search.

4 RETRIEVAL PROCESS (SEARCH STRATEGY)

Our mechanism of search is based on matching concepts for *geographic queries.* The strategy is divided into four tasks:

1.- Indexing phase.
2.- Construction of Inverted List of files.
3.- Ontology exploration.
4.- Correlation.

The first one is referred to the indexing phase, according to the context, in that case the context is: location. Then the sentence analysis recognizes spatial relations and location expressions by means of a simple heuristic, for example: ***near*** is marked as an spatial relation and ***"walking distance"*** as a location expression related to concept ***near***. Additionally, the surrounded terms to the spatial relations, also, are extracted. The result of this step is a table of terms with three attributes: the term, spatial relation and location expression, where the possible values; for the first attribute is: the word, and for the second and third one are: a true or false value according to the word.

The second step is process the table (shown in table 1) to construct an inverted list of files accord to the concepts. The result is an inverted index file shown in table 3.

The third task is exploring the ontology to find the main terms (of the extracted terms in step 1) by means of top-down exploration of ontology. The exploration starts in the root node and the node's relations represent the way to go to other concepts (related), when the term is found, the exploration finish, and then context is extracted (context = the relations and properties around the concept).

Therefore, in this step the main terms are extracted (e.g. near is the main term to "five minutes from") a term is considered as "main" when it is represented as a spatial relation in the ontology (also is a node) and has another nodes related to it. These relations and properties are searched in the table of Inverted Index files, and the documents associated to these, are returned as candidate results to the initial query.

The last task consist of correlate the set of surrounded words (*SSW)* with others *SSW's* (shown in table 2) for example *near* has a relationship with the concept *distance*, then the *SSW of near* and *SSW of distance* can have hidden relations, therefore the extraction of these is carried out. That is the point where we say that we match by concept. The representation about this process appears in Fig. 2.

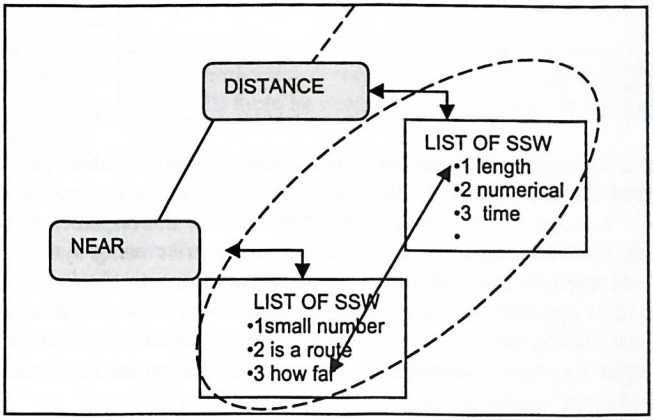


Fig. 2. Matching the Concepts NEAR and DISTANCE on ontology

The correlation process is performed using a simple relation rules, we describe two cases: the first one when documents do not contain the terms that appear in query. The firs case is described in the present section and the second case is described in Section 4.1

The first case is described considering the next situation: suppose the reference frame: "*Azteca stadium*" in a query, now in the document collection there is a *document Y* contains the reference frame: *Azteca Stadium* and we find in the ontology exploration that "*Azteca stadium*" also appears in the SSW of *distance* concept. Nevertheless, other *document X* is related to the *near* concept, because *document X* contains in the SSW the name of a street (a street near of *Azteca Stadium* but the word *near* does not appear, nor *Azteca Stadium*).

Then, the traditional retrieval will not consider the document X because there is no matching. Although in *document X* is relevant. Then, to accord to document analysis

document X and *document Y* are not related. But, both of documents are considered as candidatee to be related because in the ontology *near* and *distance* are related, and they should be appear in the final result.

The list inverted index is resumed in table 3 where the IDF (inverse document frequency), TF (term frequency) and radius search are considered to the queries related to location. After to this process will be determinate if they are correlated or not accord to the document weight. The calculation of document weight is explained in the following paragraph.

Table 3. List Inverted file used to correlate documents

Query	Document ID	Frequency proximity Word (FPW)	Reference Frame Frequency (RPF)	Resultant Frequency (RF)	Estimated Radius Search (ERS)	Final Weight Document (FWD)
1.-Near *metro station Sevilla*	001	12	2	6	1 km	60
2.-Close to *Paseo de la Reforma Avenue*	002	8	3	4	6 km	65
3.- Located inside *forest Chapultepec*	003	7	10	6	17 hectare	70

In table 3 we resume the process of correlation by using traditional inverted files but we added attributes to aid in the retrieval process, in particular in the processing queries. We can see in the first column three queries about places well known in Mexico City and very close in distance, but the queries using synonyms or other proximity expressions that requires a semantic processing to obtain relevant results, then in the table appears the main criteria to decide which documents are related, and which are not, the decision is taken by computing the values in each column to obtain a final weight for each document, this weight allows to decide if the correlation between document's terms is convenient to perform or not.

The headers of each column in table 3 are eloquent, as in traditional IR approaches the weight of document is computed based on the word occurrence number for each document. In our case we made some modifications to this process, for example the column RF is computed by considering the TF, IDF as in traditional IR theory, but adding the average between frequency proximity words (FPW) and reference frame frequency (RFF). The FPW is the number of occurrences in the document to words such as: "*near*" "*away*" "*5 minutes*", etc. RFF is the number of occurrences to places widely known such as: "*Central Bank*", "*Modern Art Museum*", etc. While that estimated radius search (ERS) column is computed based on the extension or length of the reference point, i.e. a station metro in Mexico City has an extension average of 1 km; in similar way to the other queries the value is assigned.

The final weight document column (FWD) is computed considering the ERS but restricted to the distance average to the main term, for example "*located inside*", has a related concept "*near*", then although the value RES for "*located inside*" is 17 hectare, the final average distance considered is 4 km because this is the average

value for "*near*" concept. A detail to realize is that FPW to *query 1* is higher to *query 2* and *query 3*, in the traditional approach the document 001 will be the more relevant, with our approach the document 003 is the most relevant.

That process is the last step in our strategy of search, although an additional case is included using logical inference (the user decides if the inference will be activated) that procedure is yet in test phase, but the preliminary results about initial tests are satisfactory, we describe the process in section 4.1

4.1 INFERENCE ON QUERIES

In this section, we show some queries where the relation's properties are processed to extract the semantics of a spatial relation (i.e. "*in*"). That semantics will be used to enhancement the retrieval process. Here we apply axiomatic properties such as: symmetric, transitive, reflexive, inverse to spatial relations.

The goal is performing inferences by applying first logical order, which approach is well-known but the application is oriented to geographic information retrieval. The inference process allows obtaining additional and relevant results compared with traditional approach on IR.

An example is represented in the following scenario: Suppose a user who asks for "*Rivers north of Mexico*", the semantics of *north* can be extracted with the process described below, but probably will there be concepts that do not have a direct relation with any concept.

In that case, the process can be enriched if the power of inference is used. To achieve that, we require using the results obtained from the previous queries related to the reference frames (Mexico data in that case).

These results will be stored in plain files for each query. For example, if we have the previous results for queries about *Mexico* and *Rivers*, where the documents contains expressions such as: "*USA **north of** Mexico*" and "*Wisconsin **in** USA*" then, the transitivity of "*in*" and the "*north of*" allow to infer that "*Wisconsin **North of** Mexico*", in that way we can consider candidate results that match not only by the concepts "***in***" or "***north of***", but also those whose semantics was obtained by inference. Nevertheless, additional mechanisms are required to refine and robust that phase. The table 4 shows some of the possible inferences about this particular query, according to the spatial reasoning.

Table 4. Inferences using axiomatic properties of relations

From previous results	- Wisconsin **North of** Mexico	- USA **North of** Mexico
Current query	- is Fox River **in** Wisconsin?	- is Wisconsin **in** USA ?
(Additional answers) By semantics of *in* and *North of*, we can infer that:	- Fox River **North of** Mexico	- Wisconsin **North of** Mexico

The examples shown in Table 4 describing the possible inferences to combining relations between two concepts, in particular the relations focused in direction or

location (e.g. *Northern, southern, in)* Therefore, using the transitivity property, from the previous example, is feasible infer that "*Fox River North of Mexico*", **although these concepts are not directly associated in the ontology**. With this approach we can retrieve documents using basic logical inference.

5 PRELIMINARY RESULTS

With this approach we present the following results, the document collection, at this moment, is approximately of 100 documents where the documents are web pages and pdf files, they are from several sites of internet, the criteria to search in these sites was only that documents should contain a location term.

At this moment, the relevance of results is considered as good compared with the results obtained from web search engines, the criteria of assessing is based on test using search engines based on keywords versus our approach. Nevertheless, others tests are required to evaluate and show the statistic table from overall approach.

In the Figure 3 we summarize the overall process, with an example using a query and show the steps described earlier.

Consider the following query submitted by tourist:

q= Downtown in Mexico City is near of Benito Juarez International Airport?

Some snippets from retrieved documents are shown in Fig. 3.

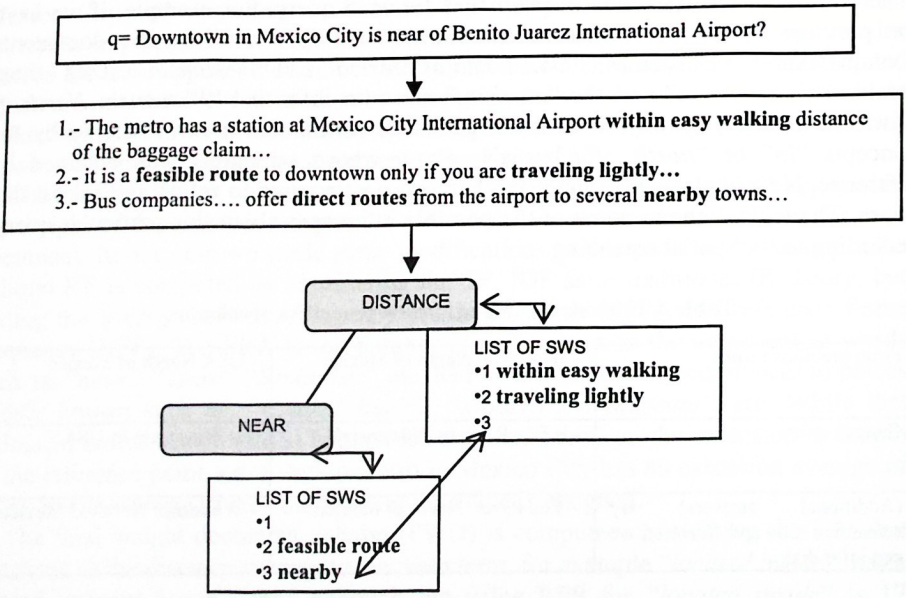


Fig. 3. Matching Concepts

In bold appears the location expression detected during the retrieval process. The first document is retrieved (using keyword match, it would not appear) in the ontology exploration phase, because the concept *near* has a relation with concept *distance*. Later, the concept *distance* has in their *SSW* list the expression: "*within easy walking*", therefore, the correlation phase return the corresponding document as a candidate result, the process continues and document appears in the final results. The case is analogous to other documents. Nevertheless additional tests are required to formalize the process.

6 CONCLUSIONS AND FUTURE WORK

This work describes an approach to performance geographic information retrieval based on location queries, the main idea is extract and process the properties and relations from geographic objects that appear in a query and documents related to them. Our approach is based on exploring ontology, it is the method used to match the concepts by relations and properties and not only by syntactical methods.

The main goal of this work is to offer more relevant results to locations query, considering spatial relations and properties from terms such as: near, in, "to five minutes from", etc. the justification is based on the fact that the actual retrieval process is performance without considering the nature of geographic object. For example: a query that contains the near term can be interpreted in different way if the topography is considered. Moreover, the primitives of representation for the surrounded terms (e.g. hotels, streets, counties) are rich in semantics, they can be the relations and properties involved in the meaning of a query.

It is not the same a query hotels near airport that hotels near main street (we require different spatial operations to obtain the results, although the spatial relation is the same in both queries).

The retrieval process is performance by means of exploring ontology and using mechanisms based on first order logic to make inferences. Moreover, using traditional retrieval information with support of ontologies can improve relevance of the results returned by traditional IR approaches and search engines.

The paper shows some examples about certain location and proximity queries. Nevertheless, it is necessary to perform additional testing to check the performance of our approach, in specific with a bigger document collection.

Our work is opposite to other techniques (e.g. query expansion) because the retrieval is performance by matching concept based on ontology exploration. Not by adding other extra keywords, based on the keyword-match. Moreover, we use the surrounding terms to query's terms as a mechanism to make a better information retrieval, because they have a lot of semantics that can be used in the retrieval process. Additionally, we use the traditional inference mechanism used in spatial reasoning. It is with the purpose of solving some location queries; without using geographic coordinates in explicit way. We define a search strategy to exploit the semantic of spatial relation, in particular those that contains a location term.

The strategy and the rules used will be formalized as part of future work. This work extends the retrieval capabilities of existing methods and proposes a method to

explore ontology as a support to GIR in conjunction with techniques used in traditional IR.

ACKNOWLEDGES

The authors of this paper wish to thank the Centre for Computing Research (CIC), General Coordination of Postgraduate Study and Research (CGPI), National Polytechnic Institute (IPN) and the Mexican National Council for Science and Technology (CONACYT) for their support.

REFERENCES

1. Wordnet: network which models the relationships between words, for example, synonyms, antonyms, hyponyms. http://www.wordnet.com

2. Egenhofer M., Interaction with Geographic Information Systems via Spatial Queries, *Journal of Visual Languages and Computing*, 1 (4): 389-413, 1990

3. Walker A.R., Pham B., Moody M.: Spatial bayesian learning algorithms for geographic information retrieval, *Proceedings of the 13th annual ACM international workshop on Geographic information systems, GIS'05*, November 4-5, 2005, Bremen, Germany

4. Delboni T. M., Borges K. A. V., Laender A. H. F.: Geographic web search based on positioning expressions, *Proceedings of the workshop on Geographic information*, 2005, Bremen, Germany

5. Maedche A., Stabb S., Stojanovic N., Studer R., Sure Y.: " A framework for developing Semantic Web Portals". *Lecture notes in Computer Science* v. 2097, 2001

6. Budak A., Sheth A., and Ramakrishnan C.:Geospatial Ontology Development and Semantic Analytics, *Handbook of Geographic Information Science*, Eds: J. P. Wilson and A. S. Fotheringham, Blackwell Publishing, 2004

7. Guarino N.: Formal ontology and information systems, *In Proceedings of the 1st International Conference,* Trento , Italy, IOS press, june 1998

8. Harding, J.: Geo-ontology Concepts and Issues, *Report of a workshop on Geo-ontology*, Ilkley UK, September, 16–17, 2003

9. Koo S., Lim S., Lee S.: Building an ontology based on hub words for information retrieval, *Proceedings of the IEEE/WIC International Conference on Web Intelligence*, (2003) 46

10. Davids Jr. C.A. ,Fonseca F.,and Borges K.: A Flexible Addresing System for approximate Geocoding. In Proceedings of V brazilian symposium on Geoinformatics, Campos do Jordão, Brazil, 2003

11. Stabb S. et al:, "Semantic Community Web Portals", 2000, *in WWW9/ computer Networks* (Special Issue: WWW9 – Proceedings of the 9th international world wide web conference, Amsterdam, the Netherlands, Maio, 15-19, 2000)

12. Heinzle F., Kcopczynsky M., and Sester M.: Spatial Data Interpretations for the intelligent access to spatial information in the internet, in Proceedings of the 21th international cartographic conference, Durban, South Africa 2003

13. Shilder F., Versley Y., and Habel C.: Extracting Spatial Information: grounding, classifying and linking spatial expressions. *In proceedings of the ACM SIGIR workshop on Geographic Information Retrieval*, Sheffield, UK, 2004

14. R. Gruber. A translation approach to portable ontology specifications. *Knowledge Acquisition*, 1993

15. Hammond, A. Sheth, and K. Kochut, Semantic Enhancement Engine: A Modular Document Enhancement Platform for Semantic Applications over Heterogeneous Content, *in Real World Semantic Web Applications*, V. Kashyap and L. Shklar, Eds., IOS Press, 2002

16. Salton G.: Automatic Text Processing: *The Transformation Analysis and Retrieval of Information by Computer*. Addison-Wesley, 1989

17. Himmelstein M.: *Local Search: The internet is the Yellow pages*, IEEE computer, 38 (2): 26-35, 2005

18. Wildöcher, A., Faurot, E., Bilhaut, F.: Multimodal indexation of contrastive structures in geographical documents. *In RIAO 2004*, Avignon, pages p.555–570, 2004

19. Torres, M.:Semantics definition to represent spatial data. *International Workshop on Semantic Processing of Spatial Data* (Geopro 2002) 2002

20. Baeza-Yates, and Ribeiro-Neto B.. Modern Information Retrieval. New York, NY: ACM Press; 1999

21. Gaizauskas, R.: An information extraction perspective on text mining: *Tasks, technologies and prototype applications. Euromap TextMining Seminar*, Sheffield, 2002

22. Jones, C.-B., Abdelmoty, A.-I., Finch, D., Fu, G., Vaid, S. The Spirit Spatial Search Engine: Architecture, Ontologies and Spatial Indexing. *Third International Conference - Geographic Information Science*, Adelphi, Usa, October 20-23, (2004) 125 – 139

23. Desai S., Knoblock C. A., Chiang Y., Desai K., Chen C.: Automatically Identifying and Georeferencing Street Maps on the Web, *The 2nd International Workshop on Geographic Information Retrieval* (GIR'05, CIKM Workshop), November 2005

24. Silveira-Chaves M., Silva M. J., Martins B.: A Geographic Knowledge Base for Semantic Web Applications. 40-54, A*CM GIR'05 workshop on GIR* (2005)

Semantic Description of Topological Relations in Spatial Databases

Miguel Martinez
Geoprocessing Laboratory-CIC- National Polytechnic Institute, Mexico City, Mexico
miguelrosales@sagitario.cic.ipn.mx,

Abstract. An approach focused on incorporating semantic content into Spatial Databases is proposed. Our methodology is based on a conceptualization of a geospatial domain restricted to recover the meaning of topological relations between geographic objects by means of concepts. Indeed, in spatial databases only a small set of topological relations is explicitly represented. While, a semantically enriched set of such relations may be required, but this sometimes can only be identified at the time when the geospatial data are displayed or analyzed by the user. Thus, we define six relations, which are obtained considering the behavior of diverse themes such as Hydrology, Land Use, Transportation Networks, and Settlements. Geospatial objects are analyzed to identify the topological relationships. We consider two analysis levels: intrinsic and extrinsic. Then, the descriptions are automatically generated in form of tuples {Oi, R, Oj}, where Oi and Oj represent a pair of geospatial objects, and R represents the concept (relation). Each tuple represents the meaning of a topological relation. For example, a highway (O1) crosses (R) a roadway (O2). The conceptual representation has some advantages with respect to the traditional approaches: the conceptualization does not depend on the data scale, geo-reference system, dimension, etc.

1 Introduction

Nowadays, the spatial databases commercially used have a little or null semantic content, great part of this content is represented implicitly in the data and requires being extracted analyzing geographic data. In general, the geospatial data have different properties that cover diverse aspect of geographical data; within these are the topological, geometrical, thematic and logic properties.

Additionally, the topological relationships between geographical objects are not explicitly represented in the spatial databases. Frequently these relationships are identified when the data are displayed or analyzed [12].

Therefore, actually it is necessary that the GIS lead the efforts on investigation to describe the spatial relationships explicitly, by means of objects conceptualization and the relationships that maintain whit other entities in some area of interest. In addition to make use of the semantic to solve problems that traditionally are dealt with numerical or classic processing, for example the interoperability into heterogenic database.

© M. Moreno-Ibarra, S. Levachkine (Eds.)
Semantics in Geoinformatics Applications
Research in Computing Science 25, 2006, pp. 35-46

When realize a topological description of spatial database in explicit way, based in concepts that represent the topological relationships, these can be stored explicitly in a spatial database. It's possible to automatically identify topological relationships analyzing different themes hat compose to a database.

With respect to analysis of relations, exits tow ways to analyzed spatial relations [15]:

- Intersection models, developed by [4] [5] [6].
- Schemas based in RCC (Region Connection Calculus), developed by [13], [14], [3], [1] and [2].

Considering these models to analyzed spatial relations, it was chosen the intersection model. Because we considerate that this model is the most adapted. This define the topological component for geographic objects, based on set point theory, which can be used to analyzed and formalized the topological relations between spatial objects. Additionally the necessaries topological components are considerate to analyzed the relations between objects that are represented with different geometrical primitives of representation, same that the RCC model does not consider.

Semantic content can be appended to a spatial database by means of tuples. These tuples can be compound by a concept that represents the topological relationships. Of this way, each tuple represent the mean of a topological database.

The context of this work, we will focus to the topology and in specific on topological database among geographic database. In order to represent semantic content in spatial databases, we use a conceptual representation of the topological relationships. Fig. 1 show the methodology that we propose to integrate semantic content to databases. And Fig. 2 show the representation conceptual of a topological description.

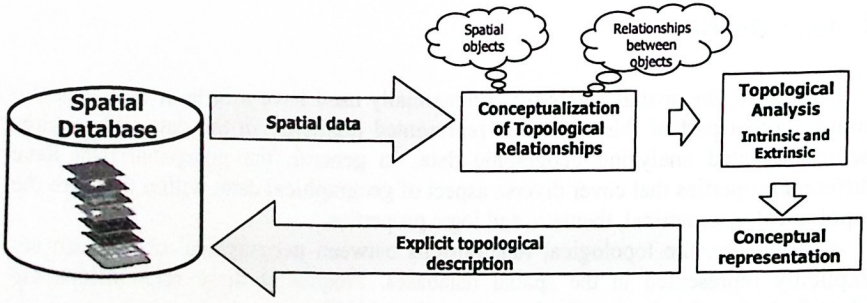


Fig. 1. Methodology proposed to integrate semantic content to spatial databases.

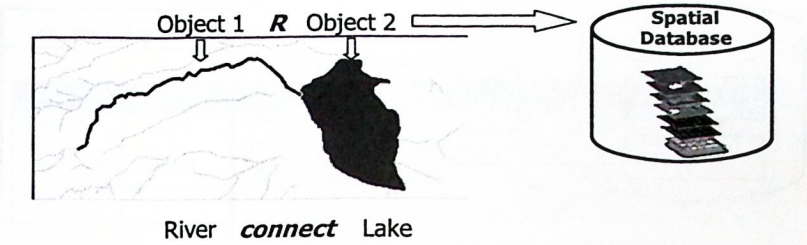


Fig. 2. Representation conceptual of a topological description.

This paper is organized as follows: Section 2 describes the schema of topological descriptor, that it is the computational system used to describe a spatial database. Section 3 describes the conceptualization of topological relationships used to describe the relations between to objects. Section 4 describes the levels of topological analysis: intrinsic and extrinsic. Sections 5 show the experimental results, and Section 6 sketches out our conclusion and future work.

2 Schema of Topological Descriptor

The design of topological descriptor consists of five principal parts. *Data source.* The system is design to allow tow source of data. These spatial data are that compose the spatial database. These data can be "shapefile" or "geofile" (property format). *Converting to geofile format.* If the source of data is "shapefile", the system should make a transformation processes y "geofile" format. *Relations analysis.* In this stage, the system makes an analysis en two levels: *intrinsic* relationships analysis and *extrinsic* relationships analysis, which are made in independent way. *Topological descriptions.* These are the result of analysis of relationships among geo-spatial objects that compose the spatial database. The descriptions are stored in dBASE files (.dbf), with this, the topologic descriptions are represented in explicit way. *Spatial database,* this component store the topological descriptions, making one table for each topologic descriptions that is obtained with that intrinsic and extrinsic analysis. These tables contain the topological descriptions and the spatial objects are stores in "geofile" format. Additionally, is possible to export the topologic descriptions for to use these in other systems. Fig. 3 shows the topological descriptor model.

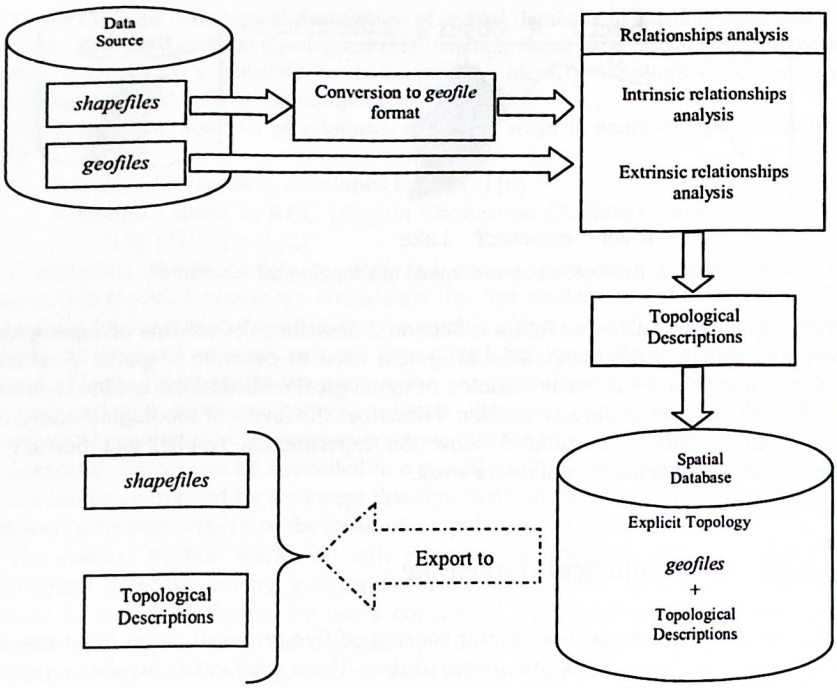


Fig. 3. Schema of Topological Descriptor.

3 Conceptualization of topological relationships

The conceptualization is performed by means of the relationships defined in 9-intersection model [6] and INEGI[1] specification [7][8][9]. We defined a set of 6 relations to describe topological descriptions between two spatial objects. These relations are based on the properties described in the topographic data dictionaries.

The description of six relations is showed in Table 1, as well as the symbol with which each relation is briefed.

[1] National Institute of Geography, Statistics and Informatics, National Mapping Agency

Table 1. Set of relations

Relations	Symbol	Description
Connect	*C*	Valid to relate linear object to another object. The initial or final node of the linear object is connected to a limit [6] of another object. For instance: A river *connect* to a prey; ; A road *connect* to another road; A highway *connect* to a population.
Share	*S*	Valid to relate area objects to area or linear objects. Pairs of this objects have common elements, except the boundary. For instance, a river that is part of the boundary of a country.
Share limit	*Sl*	Valid to relate area objects to another area objects. The only common element is the boundary. For instance, the boundary between two states.
Cross	*X*	Valid to relate linear objects to areal or linear objects. (7) part of linear object is inside of a area object; (5) two linear objects are intersected, but the flow is not shared. For instance, a highway *cross* a railroad.
Intersect	*Y*	Valid to relate pairs of linear object. This relations describe an intersection and their flow is shared. For instance, an street *intersect* with another street.
Inside	*I*	Valid to relate any kind of objects, if they are inside of an area object. For instanse, a town *inside* an state.

Connect relations is described to detail as example of as the other relations are described in [10]. This relation represents the connection between two objects. There are three variants of this relation: relations between Line/Are, Line/Line and Line/Point object. These possibilities depend on the spatial objects that are related. That is:

- When the relation is between one line object O_L and one area object O_A, we say that O_L *connect* with O_A; i.e. if some node of O_L connect with the limit of some O_A. Using the topological components [11] to define this relation, we have the following:

  $\partial O_L \cap \partial O_A \neq \emptyset$ *and* $^{\circ}O_L \cap {}^{\circ}O_A = \emptyset$ *and* $^{-}O_L \cap {}^{-}O_A \neq \emptyset$

  For example, "The river X *Connect* with the lake Y" (see Fig. 4).

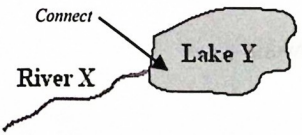


Fig. 4. The river *X Connect* with the lake *Y*.

- If the relation is between two line objects O_{L1} and O_{L2}, that is when a node of O_{L1} exist inside or in the limit of O_{L2}, we say that O_{L1} *Connect* with O_{L2}. Using the topological components to define this relation, we have the following:

  $(\partial O_{L1} \cap \partial O_{L2} \neq \emptyset$ *or* $\partial O_{L1} \cap {}^{\circ}O_{L2} \neq \emptyset)$ *and* $^{-}O_{L1} \cap {}^{-}O_{L2} \neq \emptyset$

For instance, "The river X *Connect* with the river Y" (see Fig. 5a) and "The street X *Connect* with the street Y" (see Fig. 5b).

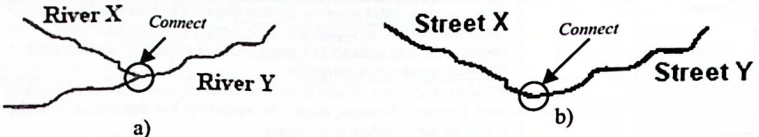


Fig. 5. a)The river X *Connect* with the river Y. b) The street X *Connect* with the street Y.

- When the relation is between some line object O_L and some point objects O_P, we say that O_L *Connect* with O_P, if O_P is in the limit of O_L. According with the Egenhofer's definition, the limit of the point object is itself. Using the topological components to define this relation, we have the following:

  $\partial O_L \cap \partial O_P \neq \varnothing$ *and* $^{-}O_L \cap {}^{-}O_P$

  For example, "The highway X *Connect* with the population Y" (see Fig. 6).

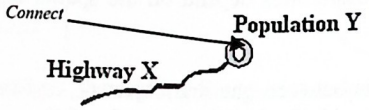


Fig. 6. The highway X *Connect* with the population Y.

The objects that they were used to compose the spatial database are classified in four thematics. The objects that were used to compose the spatial database are classified into four themes:

- Hydrology
- Land use
- Communication network
- Settlements

4 Topological analysis: intrinsic and extrinsic relationships

4.1 Intrinsic relationships

The intrinsic relationships are those that exist between objects that compose a same theme; for example, relations that exits inside Communications Networks theme. To

identify these relations, we use a diagram form each theme, like the shown in Fig. 7 to identify intrinsic relationships in Communication Networks.

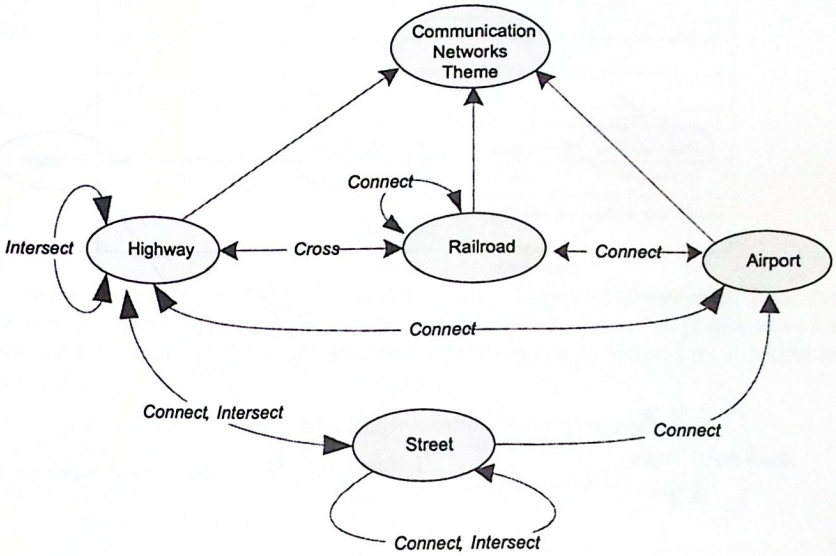


Fig. 7. Relations in Communication Networks.

4.2 Extrinsic relationships

The extrinsic relationships are those that exist between different thematics, for example, the relations between objects that belong to Hydrology and Settlement thematics. For example, in Fig. 8 is showed the relations among Hydrology and Settlement thematics.

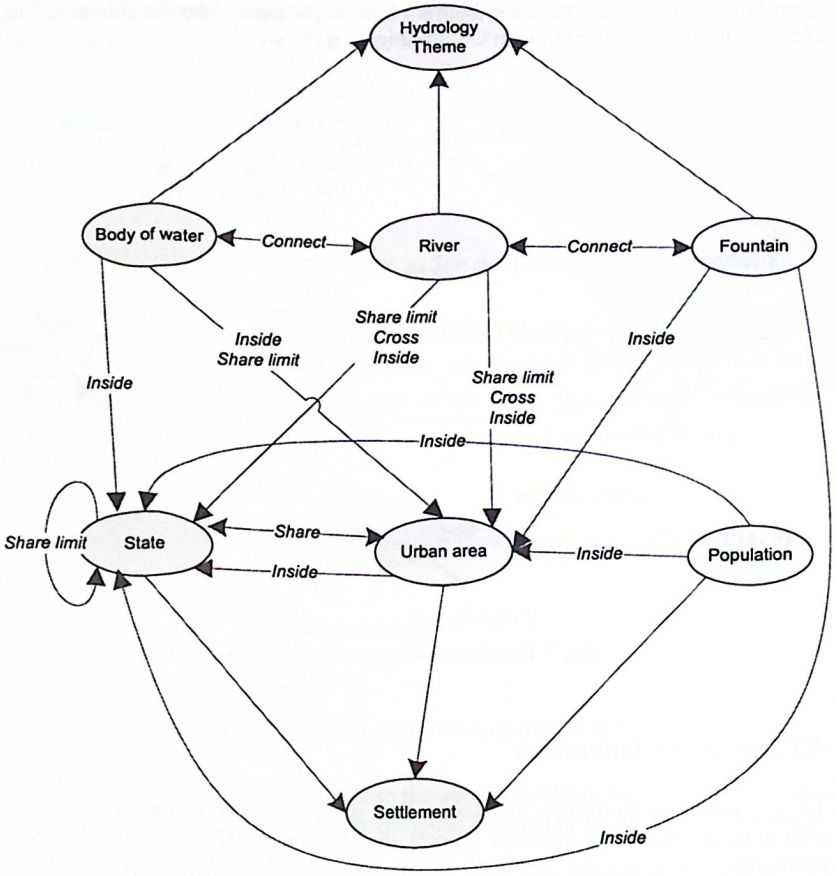


Fig. 8. Relations between Hydrology and Settlements

5 Experimental results

The results that were obtained from the topological descriptor are data stored in a table, when each table represents the descriptions of the topological relationships. These descriptions are stored in dBASE format (.dbf). Table 2 shows the attributes that compose the descriptions table.

Table 2. Attributes of descriptions table.

Attribute	Description
ID#R	Identifier existent relation between tow objects.
ID_OBJ_1	Index of first object. This Index corresponds to index of attribute table.
LAYER_BELONG_1	Data layer name of first object.
ID_OBJ_2	Index of second object. This Index corresponds to index of attribute table.
LAYER_BELONG_2	Data layer name of second object.
RELATION	Contain the symbol that identifies the relation that exists between tow objects.

To make the topological analysis was developed a library of classes that store and manage the spatial object in a proprietary format. These classes are implemented in Borland C++ Builder. The implementation of the classes is focused on working on vector data.

The functions in C++ to identify the *Inside relation* between polygons are:

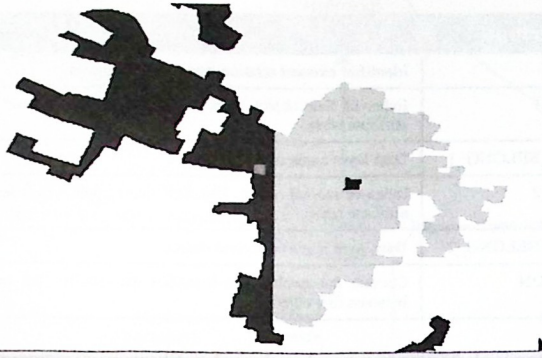
```
bool Relations::Inside(Poly *_p1,Poly *_p2){
  bool flag=false;
  if(_p2->polyInPoly(_p1))
    flag=true;
  return flag;
}

bool Poly::polyInPoly(Poly *_poly){
  bool flag=true;
  for(int i=0;i<_poly->narcs;i++)
    for(int j=0;j<_poly->arcs[i].npoints;j++)
    {
      if(pointOutPoly(&(_poly->arcs[i].point[j])))
      {
        flag=false;
        break;
      }
    }
  return flag;
}
```

The classes implement methods to obtain the basic topology (connectivity and adjacency), the topological relationships between pairs of spatial objects provide methods for computing geometrical measures, For instance, sinuosity measure of an arc.

The topological relationships are obtained using the methods *Connect()*, *Share()*, *ShareLimit()*, *Inside()*, *Cross()*, and *Intersect()*.

Fig. 9 depicts an example in which we show *Inside* relationships between two spatial objects; "Green urban area" and "Urban area" (explicitly represented in the database).



p22_uv_shp_au_shp_INT_R-I

ID#R	ID_OBJ_1	LAYER_BELO	ID_OBJ_2	LAYER_BELO	RELATION
0	0	uv_shp	3	au_shp	D
1	1	uv_shp	4	au_shp	D
2	3	uv_shp	8	au_shp	D
3	4	uv_shp	9	au_shp	D

Fig. 9. Green urban area *Inside of* Urban area.

6 Conclusions and Future Work

With this classification in thematics the analysis of relations in tow levels was carried out; reason why they were analyzed the existent relations between objects of each theme, to which it was denominated *analysis of intrinsic relations*. Afterwards it was analyzed the existent relations between thematics, to which it was called *analysis of extrinsic relations*.

The *content semantic* of relations between data is expressed by *concepts*. The conceptualization of topological relationships and concepts allow integrate a little of *semantic* to the GIS applications. The semantic content is obtained relating pairs of objects with one topological relationship. Reason why this work is one on the first attempts in this direction.

The concepts are generated using sets of geographic data. The concepts represent the interpretation of spatial data and the meaning of the relations between geo-spatial objects. With this work, we try to capture or determinate the proportions the semantic content that implicitly the spatial data contain and they do not depend on other factors, like scale or projection.

As a future work, the conceptualizations of these relations can be enhanced incrementing the number of thematics and topological relations, with which can be added new relations between objects. One interesting aspect that can be included is an analysis that depends of the context. In future will be very interest to analyze the changes in the relations that depends of context.

Is important to project that a domain conceptualization is useful to build ontologies, which represent (globally) the context of that domain, while the vocabulary of concepts and its relations describe the semantics (locally).

This descriptor can be applied to improve the results and possibly improve the performance in automatic generalization process.

Acknowledgements

The author of this paper wishes to thank to Marco Moreno-Ibarra and Miguel Torres for their support.

References

1. Cohn, A., Randell, D. and Cui, Z.: Taxonomies of logically defined qualitative spatial relations, Formal Ontology in Conceptual Analysis and Knowledge Representation, Kluwer, 1994.
2. Cohn, A., Bennett, B., Gooday, J. and Gotts, N.: Qualitative spatial representation and reasoning with the region connection calculus, GeoInformatica, 1:275-316, 1997.
3. Cui, Z., Cohn, A.G., and Randell, D.A.: Qualitative and Topological Relationships in Spatial Databases, Third Symposium on Large Spatial Databases, Lecture Notes in Computer Science N° 692, pages 296-315, Singapore, June 23-25, 1993.
4. Egenhofer, M. and Franzosa, R.: Point-Set Topological Spatial Relations, International Journal for Geographical Information Systems, 5(2): 161-174, 1991.
5. Egenhofer, M.: A Model for Detailed Binary Topological Relations, National Center for Geographical and Analysis and Department of Survarying Engineering, Department of Computer Science, University of Main, Orono, ME 04469-5711, U.S.A., 1993.
6. Egenhofer, M. and Herring, J.: Categorizing topological spatial relationships between point, line and area objects, The 9-intersections: formalism and its use for natural language spatial predicates, *Technical report* 94-1, National and Analysis, Santa Barbara, 1994.
7. Instituto Nacional de Estadística Geografía e Informática, "Diccionario de Datos Topográficos (vectorial), escala 1:250 000", 1995.
8. Instituto Nacional de Estadística Geografía e Informática, "Diccionario de Datos Topográficos (vectorial), escala 1:50 000", 1996.
9. Instituto Nacional de Estadística Geografía e Informática, "Diccionario de Datos Topográficos (vectorial), escala 1:1000 000", 1997.
10. Martínez, M.: Descriptor Topológico para Mapas Topográficos, Tesis de Maestría, México, Junio 2006.
11. Molenaar, M.: An Introduction to the Theory of Spatial Object Modelling for GIS, (Department of Geo-Informatics, International Institute for Aerospace Survey and Earth Science, Enschede, The Netherlands).
12. Mustière and Moulin, B.: What is spatial context cartographic generalisation?, Symposium on Geospatial Theory, Processing and Aplications, Symposium sur la théorie, les traitements et les applications des données Géospatiales, Ottawa, 2002.
13. Randell, D. and Cohn, A.: Modelling topological and metrical properties of physical processes, Proceedings First International Conference on the Principles of Knowledge Representation and Reasoning, Morgan Kaufmann, Los Altos, pages 55-56, 1989.

14. Randell, D., Cui, Z. and Cohn, A.: A spatial logic based on regions and connection, Proceedings Third International Conference on Knowledge Representation and Reasoning, Morgan Kaufmann, San Mateo, pages 165-176, 1992.
15. Stell, J.: Part and Complement: Fundamental Concepts in Spatial Relations, School of Computing, University of Leeds, Leeds, LS2 9JT, U. K, 2002.

Geospatial Data Processing

Using Color Information as a Whole in Image Segmentation

Rodolfo Alvarado Cervantes and Serguei Levachkine

Geoprocessing Laboratory
Centre for Computing Research (CIC) - National Polytechnic Institute (IPN)
Mexico City, Mexico
rodolfox@mexico.com sergei@cic.ipn.mx

Abstract. In this work a semiautomatic image segmentation method is presented. It uses the color information for each pixel as a whole avoiding information scattering. At present hundreds of segmentation techniques are available for color images, but most of them are monochromatic methods applied on the color components of different color spaces and then combined the results in different ways. The problem is, when the color components of a pixel are processed separately, the color information is so scattered in its components that the results are often useless for the subsequent (higher level) image processing. Our method aims at solving this problem. The improvement in quality of our segmentation technique is notorious.

1 Introduction

Image segmentation consists of partition or separation of the image in different regions, which are homogenous in some characteristic. It is an important feature of human visual perception, which appears spontaneously and naturally. It is also the first and most important task in image analysis and processing [1][3][4][8][9]; all subsequent steps like features extraction and recognition greatly depend on segmentation quality. Without a good algorithm, objects of interest cannot be recognized in slightly complex images and the whole image processing system will fail [1][2][5][9][10]. Therefore considerable effort is taken in the design of algorithms that increment the probability of a successful segmentation. At present hundreds of segmentation techniques are available for color images, but most of them are just monochromatic methods applied on the components of different color spaces and then combining the results in different ways. The problem is, when the color components of a pixel are processed separately, the color information is so scattered in its components that the results are often useless for the subsequent (higher level) image processing [4][5][6][7][9][10].

In this work a semiautomatic image segmentation method is presented, which as opposed to the previous work, uses the color information for each pixel as a whole avoiding the mentioned information scattering. In this method the three color components are combined in two parts: in the definitions of colors distances [Δ_h, Δ_s, Δ_i] and in the definition of the "Color Probability Images".

The method is primarily based on the concept of the "Color Probability", i.e. the probability that a pixel has the same color, which a user has previously selected. This color probability is calculated for every pixel with a formula developed for this

© M. Moreno-Ibarra, S. Levachkine (Eds.)
Semantics in Geoinformatics Applications
Research in Computing Science 25, 2006, pp. 49-61

purpose to build Color Probability Images (CPI's), which can, if needed or desired, be processed with any tool of mathematical morphology for gray images combining color and geometrical information in a simple way.

The improvement of the segmentation quality is quite notorious and can be appreciated on results shown in section 3 where this method is compared with one, state-of-the art commercial software, using scattering type segmentation algorithms (the only available today in literature).

2 Description of the method

Our technique can be conceptually divided in two steps:

1. The generation of color probability images (CPI's) by solving equation (1) for each pixel in the input color image. Only color information is used in this step resulting in a pixel-oriented technique.
2. The application (if necessary) of any desired morphology technique for gray images on the previously generated CPI's. In this way geometric discriminant characteristics are introduced in the segmentation process.

By the use of the concept of color probability, the color information of every pixel is processed as a whole without separating the color components. To generate a CPI is needed:

1. A color image in RGB format.
2. A set of pixels forming a sample of the color desired

From the sample we calculate statistical values according to a HSI modified model (section 2.6). To obtain a CPI we calculate for every pixel in the color image the color probability CP(i,j):

$$\mathbf{CP(i,j)} = e^{\left(\frac{-a_1\Delta_h^2}{2\sigma_h^2}\right)} * e^{\left(\frac{-a_2\Delta_s^2}{2\sigma_s^2}\right)} * e^{\left(\frac{-a_3\Delta^2 i}{2\sigma_i^2}\right)} \quad (1)$$

e = 2.7182...
Δ_h = Hue_distance_Δ_h (hue(i,j), average_hue)
Δ_s = saturation_distance_Δ_s (saturation(i,j), average_saturation)
Δ_i = intensity_distance_Δ_i (intensity (i,j),average_intensity
σ_h = Hue standard deviation
σ_s = Saturation standard deviation
σ_i = Intensity standard deviation
$[a_1\ a_2\ a_3]$ = Mask of refinement, a_i can be 0 or 1

Some modifications on standard HSI color space were necessary in order to create a consistent model to represent color and color centroids:

1. *Representation of hue.* Instead of standard representation of hue as an angle between [0 360], hue is represented as a normalized vector in R^2 (with magnitude 1 or 0). This representation has at least 3 advantages compared to an angle [0 360]: a) The existing discontinuity in 360 and 0 degrees is eliminated. b) Hue average of a group of pixels can be understood as the resulting angle of the vector addition of the pixels in the chromatic region of the sample, giving a simple manner to calculate the average. c) Setting magnitude to 0 or 1 works as a flag intended for distinction between chromatic or achromatic regions.

2. *Separation of chromatic and achromatic regions of the HSI space.* We use a separation of the region (according to [1]) with possibility of refining with the $[a_1\ a_2\ a_3]$ mask (in equation (1)) in order to calculate the hue average and Δ_h. Once calculated Δ_h, Δ_s and Δ_i this distinction is not longer necessary because in the formulation of CP (equation (1)) all the cases of color comparison between zones are accounted and maintain consistency.

2.1 Hue average calculation

In order to obtain a value of hue, which represents the average of several pixels of a sample, we take advantage of the vector representation in R^2; the vectors that represent the hue values of individual pixels are combined using vector addition. From the resulting vector we obtain two significant values used in the algorithm: average hue and relative saturation. Thus Hm is calculated in this manner:

1. For every pixel in the sample the following R^3 to R^2 transformation is applied:

$$VR(P)= \begin{cases} \begin{bmatrix} 1 & -\cos(\pi/3) & -\cos(\pi/3) \\ 0 & \sin(\pi/3) & -\sin(\pi/3) \end{bmatrix} * \begin{bmatrix} R \\ G \\ B \end{bmatrix} = \begin{bmatrix} x \\ y \end{bmatrix} & \text{If } P \notin G \\ \begin{bmatrix} 0 \\ 0 \end{bmatrix} & \text{If } P \in G \end{cases}$$

G represents the achromatic zone in the HSI space. [RGB] are the color components of the pixel in the RGB color space

2. Equivalent to the following pseudocode is executed

```
Vector.x = 0;
Vector.y = 0;              // initialize vector
For (i = 1; i < = n; i++)          // for every pixel in
the sample do
{
    Vector.x = Vector.x +   VR(i).x;
```

```
Vector.y = Vector.y +  VR(i).y;}
```

In this code we have a vector, which accumulates the vector additions as index i increments. Each of the vectors being added corresponds to the previous R^3 to R^2 transformation for every pixel in the sample made in step 1.

3. The angle respect to the X-axis that is obtained corresponds to average: Hm = angle (Vector,0).

4. Relative saturation Sr is calculated: Sr = Magnitude(Vs) / Magnitude(V1) + Magnitude(V2)

2.2 Achromatic region G

The achromatic zone **G** is the region in the HSI color space where no hue is perceived. That means color is perceived as gray levels because color saturation is very low or intensity is either too low or too high.

Given the three-dimensional HSI color space, we define the achromatic region **G** as the union of the points inside the cylinder defined by Saturation < threshold_1 and the two cones Intensity < threshold_2 and Intensity > threshold_3. Pixels inside this region are perceived as gray levels (Fig. 1).

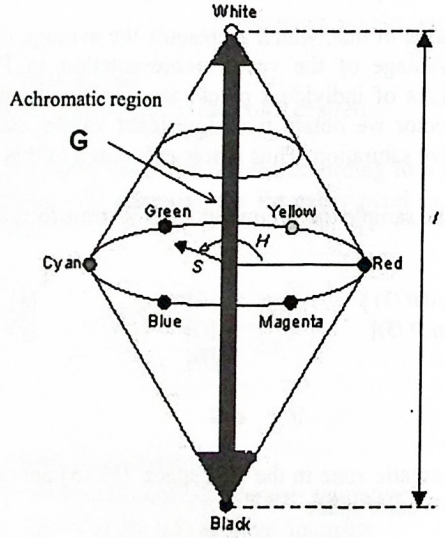


Fig. 1. Achromatic region **G**

2.3 Hue distance Δ_h

Using the vector representation of Hue, we define hue distance Δ_h for two pixels P_1 and P_2 as follows:

$\Delta_h (P_1,P_2) =$| K*abs(arcos[dot_product(V_1,V_2)/mag(V_1)*mag(V_2)]) If P_1 and $P_2 \notin G$
| 0 If P_1 or $P_2 \in G$.

abs () = absolute value
arcos () = arc-cosine
mag () = magnitude of vectors
dot_product () = vector dot product

K is a normalizing real factor to force $\Delta_h \in [0,1]$.
G is the achromatic region
V_1 y V_2 are the vectors in R^2 calculated with the transformation on P_1 and P_2:

Vm (P)=

$$\{[1 \quad -\cos(\Pi/3) \quad -\cos(\Pi/3)] \quad * \quad [R] \quad = \quad [x] \quad \text{If } P \notin G$$
$$[0 \quad \sin(\Pi/3) \quad -\sin(\Pi/3)] \qquad [G] \qquad [y]$$
$$[B]$$
$$[0]$$
$$[0] \qquad \text{If } P \in G\}$$

2.4 Saturation distance Δ_s and intensity distance Δ_i

Saturation distance and intensity distance between two pixels is defined as:

Δ_s = abs [saturation(P_1)-saturation(P_2)] and Δ_i = abs [intensity(P_1)-intensity(P_2)]

2.6 Statistical values for groups of pixels

The statistical values needed in equation (1) are calculated as follows:

$$\text{Saturation average} = \frac{\sum_{i=0}^{n} saturation(i)}{n}; \tag{1}$$

$$\text{Intensity average} = \frac{\sum_{i=0}^{n} intensity(i)}{n} \tag{2}$$

$$\text{Hue standard deviation } \sigma_h = \sqrt{\frac{\sum_{i=1}^{n} \Delta^2_h(i)}{n-1}}; \tag{3}$$

$$\text{Saturation standard deviation } \sigma_s = \sqrt{\frac{\sum_{i=1}^{n} \Delta^2_s(i)}{n-1}} \tag{4}$$

$$\text{Intensity standard deviation } \sigma_i = \sqrt{\frac{\sum_{i=1}^{n} \Delta^2_i(i)}{n-1}} \quad (5)$$

n is the number of pixels of the sample.
Δ_h = hue distance_Δ_h (hue(i,j), hue_average)
Δ_s = saturation_distance_Δ_s (saturation(i,j), saturation_average)
Δ_i = intensity_distance_Δ_i (intensity(i,j), intensity_average)

2.7 Use of mathematical morphology on CPI images

The CPI image represents the image pixels that has the maximal probability of belong to the previously selected color. As this image is a gray image, can be treated with any tool of mathematic morphology for gray images. Filters, operators, thresholds, etc can be applied directly to the CPI strengthening the segmentation as geometrical characteristics are introduced. The common intensity image can be processed too as a complementary information source.

3 Experimental results

Our segmentation method was applied to a section of a topographic map (See Fig. 2)

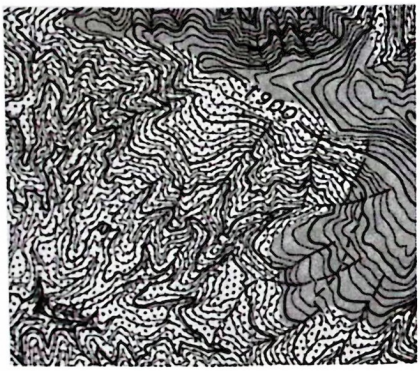


Fig. 2. Section of a topographic map

The thematic layers contained in the image are:

1. River layer (blue lines)
2. Gross brown isolines
3. Thin light brown isolines
4. Green area

1. River layer (blue lines)
2. Gross brown isolines
3. Thin light brown isolines
4. Green area
5. Green spots

To segment the river layer we took a pixel sample, which looks representative at a simple view (See Fig. 3)

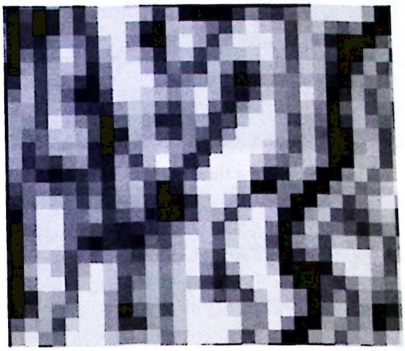


Fig. 3. Sample of blue pixels of the river layer

2. Using this sample we calculated the color centroid and deviation:

Color centroid $\{Hc \in [0,2\pi], Sc \in [0,1], Ic \in [0,255]\} = (3.56, 0.4126, 102)$
$\sigma_{ti} = 0.0392$; $\sigma_{si} = 0.0884$; $\sigma_{ii} = 0.0382$
3. We generated the corresponding CPI (Fig. 4)

Fig. 4. CPI of blue pixels

4. We applied a simple threshold to this image (Fig 5) to segment the layer. The range for values of thresholds is very wide (similar results were obtained for range [10 150]) we choose value equal to 30:

Fig. 5. Result for threshold of 30 on CPI

We observed that even with a coarse selection of the pixel sample, we obtain a good separation of this layer. Several different pixel samples were taken variating the number of pixels from 3 to 10; all of them gave very good results.

Looking at Fig. 2 we can observe that color is a discriminant characteristic of the river layer. That means no other layer has a color, which a human can confuse with this river layer. That is why we can segment it so easily only with the threshold (with a wide range) of corresponding CPI. This doesn't happen with the other layers; so additional morphological processing is needed. In some cases already extracted layers are necessary to extract some others (case of Fig. 12).

Fig 6 can be considered very acceptable as input for recognition, but we can refine it in this way: Apply dilatation followed by elimination of small disconnected areas. The result in shown in Fig 7.

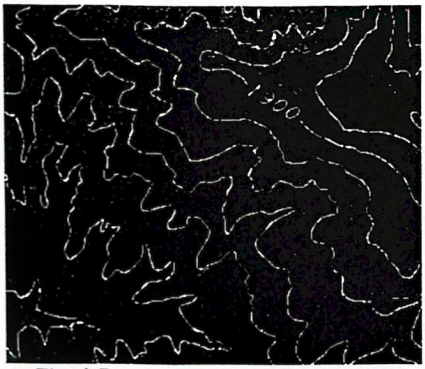


Fig. 6. Result for threshold of 80 on CPI

Fig. 7. Segmentation of gross brown isolines

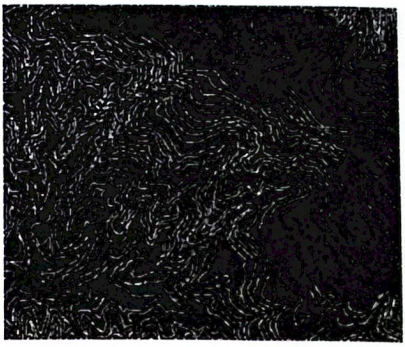


Fig. 8. CPI image of thin light brown isolines

To improve the quality of Fig. 9 we subtract the dilatation of Fig 7 to eliminate any confusion with that layer. The result is shown in Fig 10. For the green area and the green spots we present the final results (Fig. 11 and Fig.12).

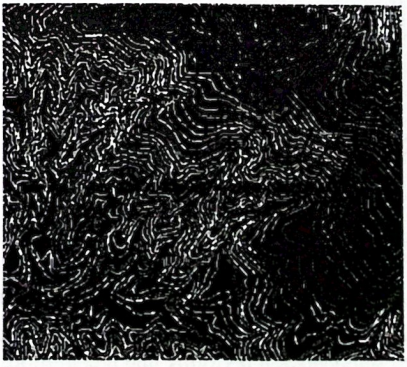


Fig. 9. Result for threshold of 150

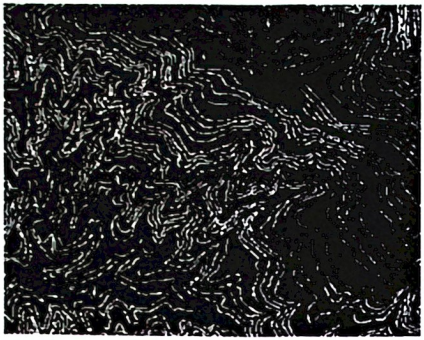


Fig. 10. Segmentation of thin light brown isolines

Fig. 11. Segmentation of green area

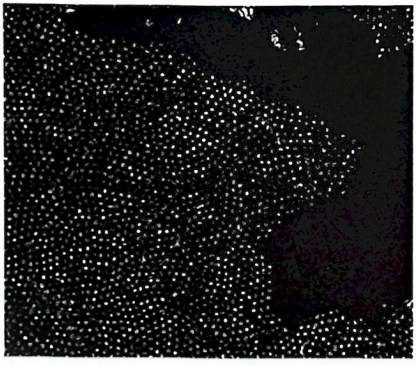


Fig. 12. Segmentation of green spots

Now we process the same map using the software R2V from Able Corp. This software additionally to segmentation performs line recognition, so not all the points shown are obtained from the segmentation step (See Fig 13, 14, 15, 16, 17).

Fig. 13. Segmentation of river layer

Fig. 14. Segmentation of gross brown isolines

Fig. 15. Segmentation o thin light brown isolines

Fig. 16. Segmentation of green spots

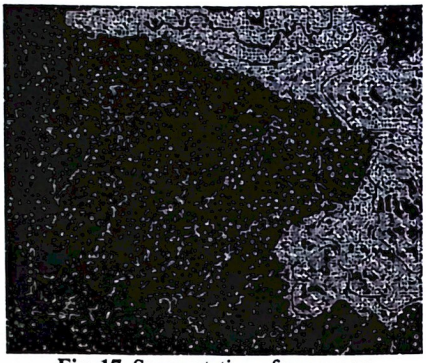


Fig. 17. Segmentation of green area

After comparing the results of our system with those from R2V, we can observe that in R2V there are a large number of pixels that should appear in the segmentation of the layer but they do not. The worst case is in Fig. 13 that is practically empty. We can also observe in R2V that there is a big amount of confusion of pixels from other layers for example in Fig. 14 segmentation of gross brown isolines, there appears many pixels from the green area that should not, and in Fig. 15 appears pixels from all layers.

4 Conclusions

From an analysis of the results we can state that the method exposed in this paper offers a useful and efficient alternative for the segmentation of color images. We cannot say the same about R2V because none of the results is even close in quality to the thresholded CPI's. This gives us a clue that in R2V the color information is not

well used as in a CPI. From experiments in many different maps we can mention some advantages of our technique:

1. *Simplicity.* Required steps to obtain a good segmentation of the layer of interest are usually simple and repetitive. If color is a discriminative characteristic of the layer of interest, only threshold of the CPI is needed to obtain a good segmentation.
2. *Very low computational complexity.* The generation of a CPI only requires solving equation (1) for every pixel. Thus the complexity is linear with respect to the number of pixels of the source image.
3. *Layer separation by color.* Generally other methods extract punctual and linear objects together, making necessary much additional work to separate those layers
4. *Good segmentation of layers using color only.* From many experiments we observed that a good percentage of layers were obtained from threshold of CPI only.

References

1. Plataniotis, K.N., A.N. Venetsanopoulos: Color Image Processing and Applications, 1st Edition Germany, Springer, 354p
2. Alvarado Cervantes, Rodolfo, Master Thesis: Segmentación de patrones lineales, topológicamente diferentes, mediante agrupamientos en el espacio de color HSI, Centre for Computing Research (CIC) - National Polytechnic Institute (IPN), Mexico 2006
3. Serguei Levachkine: Raster to Vector Conversion of Color Cartographic Maps. GREC 2003: 50-62
4. Serguei Levachkine, Efrén González-Gómez, Miguel Torres, Marco Moreno, Rolando Quintero: Knowledge-Based System for Color Maps Recognition. KES (1) 2005: 297-303
5. Angulo, Jesús, Serra Jean : Mathematical morphology in color spaces applied to the analysis of cartographic images, GEOPRO 2003: 59-66
6. Efrén González-Gómez, Serguei Levachkine: Color Cartographic Pattern Recognition Using the Coarse to Fine Scale Method. CIARP 2004: 533-540
7. Serguei Levachkine, Miguel Torres, Marco Moreno, Rolando Quintero: Knowledge-Based Method to Recognize Objects in Geo-Images. KES 2004: 718-725
8. H.D. Cheng, H.D. Jiang, Y. Sun & Jingli Wang: Color image segmentation: advances and prospects, Pattern Recognition 2001, 34(12), 2259-2281
9. Hanbury, Allan; Serra, Jean: A 3D-polar Coordinate Colour Representation Suitable for Image Analysis, Technical Report PRIP-TR-77 Austria 2002
10. Serguei Levachkine, Miguel Torres, Marco Moreno, Rolando Quintero: Simultaneous Segmentation-Recognition-Vectorization of Meaningful Geographical Objects in Geo-Images. CIARP 2003: 635-642

Conceptual Representation of Digital Elevation Models

Rolando Quintero and Virginia Villegas

Laboratorio de Geoprocesamiento- Centro de Investigación en Computación - IPN,
Av. Juan de Dios Bátiz, Esquina con Miguel Othón de Mendizábal,
México, D.F., 07738. México
{quintero, villegas}@cic.ipn.mx

Abstract. In this work, we present a method for the conceptual representation of digital elevation models. This method has four stages, the first stage is focused on the conceptualization of objects and its intrinsic characteristics, the result of this stage is an ontology. The second stage is the mapping from conceptual characteristics to numeric values, for this purpose we will use fuzzy sets. The result of this stage is a set of bands that represent the behavior of the model. The third stage is based on carrying out the classification of the digital elevation model considering the bands obtained in previous stage; we use a semantic classification algorithm, which has been developed for this task. The last stage compiles the results of the classification and puts them in a data representation according to the ontology obtained in the first stage.

1 Introduction

Digital Elevation Models (DEMs) are performing an important role in several fields of Geographic Information Systems (GIS), including the environmental science, risk prevention, engineering and so on. The geometric characteristics of DEMs (resolution, coordinates, rows and columns numbers, etc.) describe the thematic aspects of the terrain, which are represented by qualities. Also, the use of the geometric shape terrain to analyze its distribution and the concentration of certain geospatial objects has been incorporated to DEMs. These representations have been traditionally applied to distinguish the dividing lines of water, drainage and other groups of the terrain objects.

The use of elevation models allow us to review six basic foundations: the relation between the terrain characteristics and the geomorphologic process, analysis of scales, analysis of changes in the surface, analysis of flow or surface movement, visualization approaches and topographic models [3].

DEM processing has been treated by using a numeric approaches [4][13]. In the present work we attempt to make a processing from the semantic point of view. We propose a procedure to represent DEMs in a semantic way. This process is depicted in Fig. 1.

© M. Moreno-Ibarra, S. Levachkine (Eds.)
Semantics in Geoinformatics Applications
Research in Computing Science 25, 2006, pp. 63-71

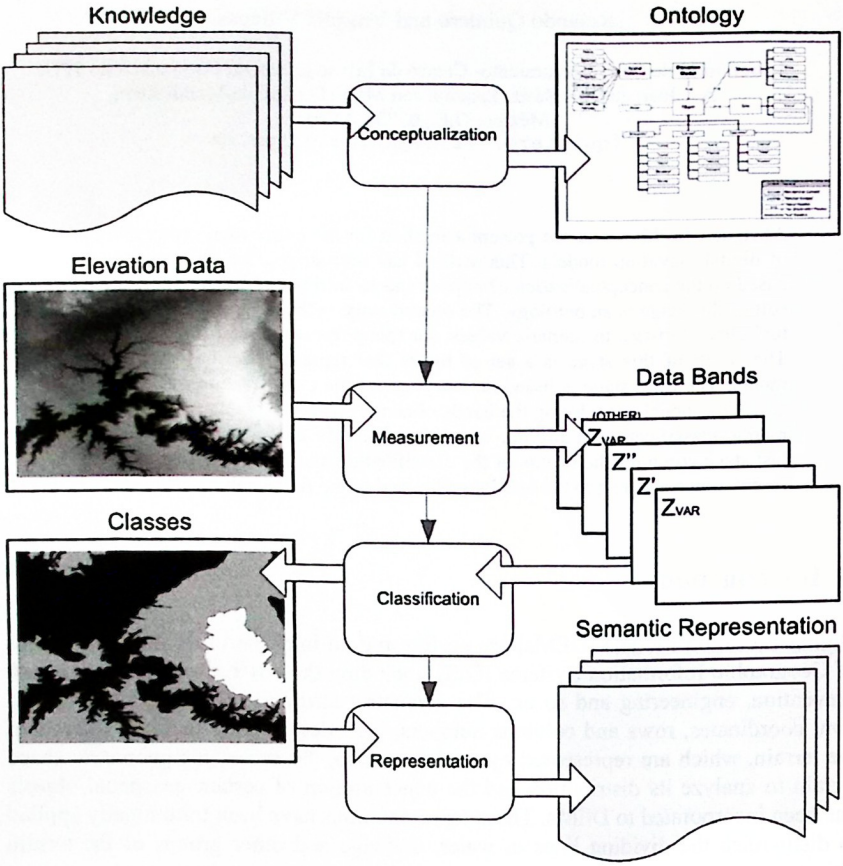


Fig. 1. Procedure proposed for conceptual representation

The spatial semantics is based on the description of the intrinsic properties of geospatial objects. These properties depend on the organization or the *status* of the object. For instance, the width and area of a polygon can provide a description. With this description, it is possible to generate specific standards based on the characteristics of the geospatial representation primitives (lines, points and areas), which define the behavior and relations between geographic objects.

Up-to-date, many analysis involved *quantitative characteristics*, therefore we propose to make a conceptualization focused on the *qualitative properties* of the geospatial objects and their relations in a general analysis. The characteristics that we consider to generate the conceptualization of the model are the follows: slope, surface, extension ruggedness and altitude.

The rest of the paper is organized as follows. Section 2 describes the conceptualization process. The measurement of geographic characteristics is presented in Section

3. In Section 4, we point out the semantic classification algorithm and the conceptual representation. Finally, Section 5 sketches out the conclusions of the research.

2 Conceptualization of spatial data in elevation models

The conceptualization of the characteristics on digital elevation models (elevations, depressions, plains, etc.) is the first stage of the method. It is done by taking the definitions, given in the Royal Spanish Academy [9], about the characteristics mentioned in section 1. Also, we use definitions given by National Institute of Statistics, Geography and Informatics of Mexico (INEGI) [5] related to the same characteristics. The analysis of those definitions provides as result ontology. This ontology represents the concepts that define the elevation model's characteristics, as well as the relations between those concepts. So, in order to represent the conceptualization of geographic characteristics[1], we will use ontology in the way described in [11]. In this work, we proposed to describe the ontology using two types of concepts (terminal and non-terminal ones) and only two types of relations (*"is-a"* for specialize concepts and *"has"* for aggregate concepts).

As it has been said, we define two types of concepts (C) in the ontology: *terminal* (C_T) and *non-terminal* (C_N) *concepts*. The first ones are concepts that do not use other concepts for defining their meaning (they are defined by "simple values"). The meaning of non-terminal concepts is defined by other concepts (terminal or non-terminal) (Eqn. 2).

$$C = C_N \bigcup C_T \tag{2}$$

Each concept has a set of aspects. They are properties and relations between geographic entities[2]. In the following, we shall use the term "relation" to denote unary relations/properties as [2]. From this point of view, all aspects of a terminal concept are simple, e.g. the type of all aspects that belongs to the set of primitive types (T_P), as shown in Eqn. 3.

$$\begin{aligned} &T_P = \{number, character, string, enumeration, struct\}, \\ &A = \{a_i \mid type(a_i) \in T_P\}, \end{aligned} \tag{3}$$

where T_P is the set of primitive types; A is the set of aspects.

Then, the set of *terminal concepts* is defined by Eqn. 4.

$$C_T = \{c(a_1, a_2, ..., a_n) \ni a_i \in A,\ i = 1,..,n\} \tag{4}$$

In the same way, the *non-terminal concepts* have at least one aspect that does not belong to T_P. It is denoted by Eqn. 5.

$$C_N = \{c(a_1, a_2, ..., a_n) \ni \exists a_i \notin A\} \tag{5}$$

---

[1] Geographic objects.

[2] Attributive data

where c is a concept.

Finally, the set of relations R is defined by the pairs that are associated to Γ and Φ, where Γ and Φ are non-reflexive, non-symmetric, and transitive relations (Eqn. 6).

$$R = R_\Gamma \cup R_\Phi = \{(a,b) \mid a\Gamma b,\ a \in C_N,\ b \in C\} \cup \{(a,b) \mid a\Phi b,\ a \in C_N,\ b \in C\} \quad (6)$$

Fig. 2 shows a fragment of the ontology used for the elevation model's conceptualization. In the figure, the concepts denoted by *"..."* (Three points) represents 'other' concepts that are presented in the same category of their "brothers" in the ontology. For instance, look at the concept *"depresión"* and its children concepts (*"cuenca"*, *"valle"* and *"collado"*), it is denoted by *"..."* that other *"depresión"* children can exist (*"cañada"*, *"cañón"*, *"barranca"*, etc.). In the ontology we must define all required concepts for describing geospatial data in elevation models, according to INEGI [5].

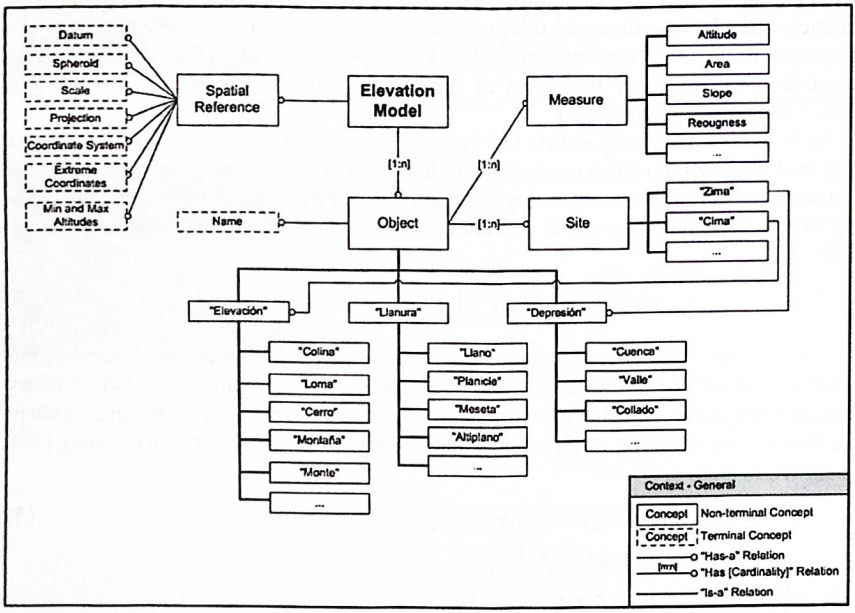


Fig. 2. Fragment of the ontology used for the conceptualization of geospatial data in elevation models. The ontology consists of a set of concepts and a set of relations. There are two types of concepts (terminal and non-terminal) as well as two types of relations (*has* and *is-a*).

3 Measurement of geographic characteristics

Once we have made the conceptualization, the second stage consists of assigning metrics to the concepts in the ontology. These metrics are ranges or procedures to

obtain numeric values from the elevation data. For instance, the concept *"cima"* (highest point of mountains, mount and hills) can be found numerically from the elevation model. Other type of values determine the difference between concepts, as in the case of the concepts *"colina"* (natural land elevation, lower than a mountain) and *"monte"* (natural land elevation, higher than a hill), In this example, the boundary between the concepts is not clear; we only know that a *mount* is 'higher' than a *hill*. In cases like the last one, we will use the INEGI definitions in order to have a hint about the boundary. Anyway, we will consider that this boundary is always diffuse.

On the other hand, an elevation model is fundamentally a discrete function Z of world coordinates, represented by means a matrix containing the elevation data (Eqn. 7).

$$Z(m), \tag{7}$$

Also, some other metrics can be determined using the Z function as a base, for instance, the slope (Eqn. 8), curvature (Eqn. 9), the variance of altitude (Eqn. 10), among others.

$$Z' = \sqrt{\left(\frac{\Delta Z}{\Delta x}\right)^2 + \left(\frac{\Delta Z}{\Delta y}\right)^2}, \tag{8}$$

$$Z_x'' = \frac{\delta^2 Z}{\delta x^2}, Z_x'' = \frac{\delta^2 Z}{\delta y^2}, \tag{9}$$

$$Z_{VAR} = \frac{1}{MN}\sum_{i=1}^{M}\left(\sum_{j=1}^{N}\left(Z(i,j) - \bar{Z}\right)^2\right), \tag{10}$$

With these metrics we obtain 'pages' or 'data bands' that will be used in the classification process.

By using the characteristics modeled in the conceptualization stage, we can obtain the classification of vectors ω, that will be used later in the process. It is necessary to recall that the conceptualization stage provides a set of concepts (within ontology) with diffuse properties. For example, as we previously mentioned, in the case of the concepts *"colina"* and *"monte"*, the boundary between the concepts is not clear; we only know that a mount is 'higher' than a hill. In this case, we can distinguish with the statements 'A *"colina"* (hill) has low altitude and small area' and 'A *"monte"* (mount) has medium altitude and large area'. This carries us to define some fuzzy sets in order to map numeric values to conceptual ones (see Fig. 3).

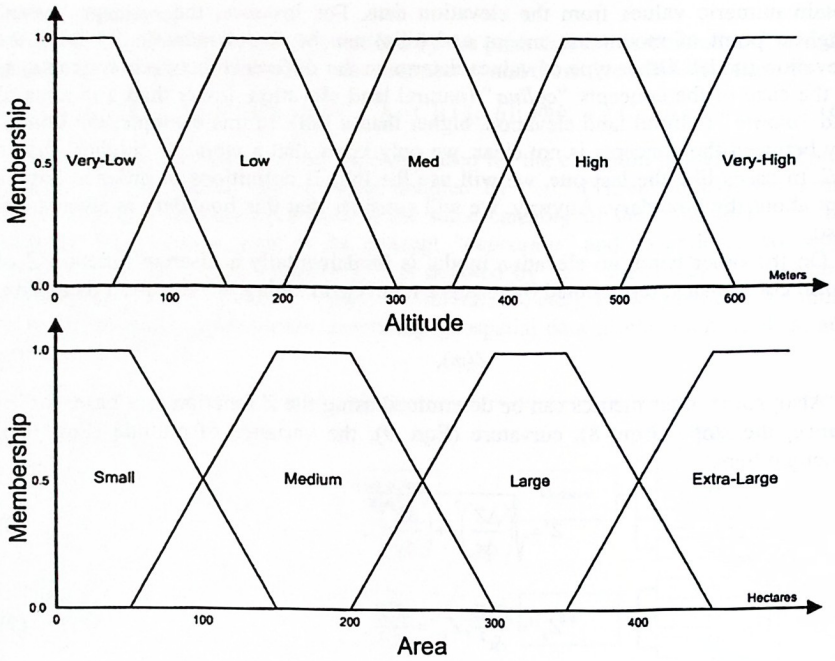


Fig. 3. Fuzzy sets obtained from the conceptualization of altitude and area of a geographic object.

4 Classification and representation of geometric characteristics

The third stage is the classification of the digital elevation model, according to the metrics that have been obtained of the conceptualization process[3]. The metrics that can be made on the pixels are the follows: altitude, slope, planar curvature and ruggedness [3]. Although in [7] is presented other measurements that can be used to this purpose. With these measurements to each pixel is obtained a set of "pages", which describes the behavior of the DEM according to each metric. For the classification, it is not necessary to have training stages nor computing the characteristic vectors, since the conceptualization provides the classes for the classification and the characteristic vectors. With this information, it is possible to apply some classification algorithm [1][6][8]. In [10] is sketched out an algorithm for semantic supervised segmentation to classify multispectral geo-images.

[3] It is important to mention, there are not so much metrics that can be made on DEMs. It can be a constraint for this procedure. However, future works are oriented to incorporate information of different sources associated to DEMs such as satellite images.

The result of the conceptualization stage is the set of semantic characteristic vectors Ω of the classification. From the Ω set, it is necessary to compute the mean and covariance matrix for each $\omega_i \in \Omega$. By attempting to determine the class or category for each pixel at a location x, it is necessary that each pixel contains a conditional probability, denoted by Eqn. 11:

$$p(\omega_i \mid x), i = 1,\ldots, M \tag{11}$$

In [12], we describe the supervised clustering method for more details. Therefore, Bayes theorem provides potential means of converting knowledge of predictive correlations. The constraint (Eqn. 12) is used in the classification algorithm, since the $p(x)$ are known by training data, we assume that it is conceivable that the $p(\omega_i)$ is the same for each ω_i, due to this, the comparison is $p(x \mid \omega_i) > p(x \mid \omega_j)$.

$$x \in \omega_i \text{ if } p(x \mid \omega_i)p(\omega_i) > p(x \mid \omega_i)p(\omega_j) \text{ for all } j \neq i \tag{12}$$

In this analysis, we assume that the classes have multidimensional normal distributions and each pixel can assign a probability of being a member of each class. After computing the probabilities of a pixel being in each of the available classes, we assign the class with the highest probability. The algorithm consists of the following steps:

[Step 1]. Given the number of classes ω_i by means of the conceptualization process, compute the maximum likelihood distribution and the covariance matrix for each class ω_i.

[Step 2]. For each image pixel, determine its semantic vector.

[Step 3]. Compute the probability of vector s to know if it belongs to each class ω_i.

[Step 4]. Obtain the coordinates (x, y) of the pixel, if the constraint (see Eqn. 13) is accomplished, then the pixel belongs to the class ω_i.

$$p(x \mid \omega_i)p(\omega_i) > p(x \mid \omega_j)p(\omega_j) \text{ for all } j \neq i \tag{13}$$

[Step 5]. Repeat from the step 2 until all pixels in the elevation model can be classified.

Generally, the classification result emits *noise*, since this process is made at pixel level. As an additional step for the classification stage, it is necessary to determine "consolidated regions". This can be performed in several ways, the best option but more complex is the "semantic consolidation". However, it is necessary to conceptualize the problem to define rules by means of case study for its consolidation. For instance, the rule: *"if a decline is surrounded by a mountain, then the decline must be absorbed by the mountain"*. In this rule there are features already conceptualized, such as the *decline* and the *mountain*, but there are not yet other features conceptualized, i.e., *surrounded* and *absorbed*. An alternative, it is to use algorithms oriented to growth of regions, in which small regions are merged by big regions, until a boundary is defined among these regions.

The last stage consists of compiling the results about the classification to put in a conceptual way into a description[4]. Future works are focused on linking a toponym database to label regions according to their names. At this moment, the description is restricted to specify the existence of geographic characteristics and some important attributes such as surface, altitude mean, borders, etc.

In this research, we have not considered the conceptualization of topological relations that are presented in the geographic characteristics. However, we have envisaged incorporating these relations in a short time, because the topological relations are indispensable for consolidating semantics of regions, as well as for making a more descriptive representation.

By using this method, we attempt to approximate the semantics that contain DEMS, which will be useful to evaluate the semantic similitude between different models. According to this point, the evaluation uses concepts that are organized in a hierarchy[5]. With this focus, we attempt to use the qualitative measurements instead of quantitative measurements in order to evaluate the similitude and measure the distance between concepts stored in the conceptual representation.

5 Conclusions

In this work, we have presented a method to make a semantic representation of elevation models. The method has four stages: conceptualization, measurement, classification and representation. In the first stage, we propose the conceptualization of the objects that can be found within an elevation model, as well as the characteristics that define these objects. The result of this stage is the ontology that conceptualizes the domain. The second stage consists of mapping the conceptual characteristics to numeric values. We propose the use of fuzzy sets in order to make this mapping. The result of this stage is a set of data bands that represents the behavior of elevation model according to a specific characteristic. In the third stage, it is carried out the classification of the elevation model. In the classification process, the data bands obtained in the second stage are used as input for the semantic classification algorithm. The last stage takes the classification results, and puts them into a semantic representation according to the ontology obtained in first stage.

This work is in progress, so we have a lot of tasks to do. In a general way, it is necessary to grow the ontology with the conceptualization of topologic relations between geographic objects, as well as the metrics conceptualization. Also, we must set up the real values of fuzzy sets, introduced in Section 3, in order to make the values available to the classification process. Finally, we must implement the method[6], as well as the necessary tests.

---

[4] List of objects (and its attributes) found in model

[5] The term of hierarchy is defined as a data structure that stores concepts that are related by one relationship, in which the partitions are completed to represent the knowledge about a certain context.

[6] We have already implemented the semantic classification algorithm.

Acknowledgments

The authors of this paper wish to thank the CIC, SIP, IPN and CONACYT for their support.

References

1. Bandyopadhyay, S.: Satellite image classification using genetically guided fuzzy clustering with spatial information. *International Journal of Remote Sensing*, Taylor & Francis, Vol. 26, No. 3, (2005), pp. 579-593.
2. Berendt, B., Barkowsky, T., Freksa, C. and Kelter, S. I. E.: Spatial Cognition - An Interdisciplinary Approach to Representing and Processing Spatial Knowledge. *Spatial Representation with Aspect Maps*, Springer-Verlag, (1998) 157-175.
3. Etzelmüller, B and J.R. Sulebak: Developments in the use of digital elevation models in periglacial geomorphology and glaciology. *Physische Geographie*, Vol. 41, Jahrestagung SGmG Winterthur, (2000).
4. Evans, I.S.: General geomorphometry, in Goudie, A.S. (ed.), *Geomorphological Techniques*, (1981), pp.31-37.
5. Intituto Nacional de Estadística, Geografía e Informática de México (http://www.inegi.gob.mx).
6. Jianwen, Ma., Bagan H.: Land-use classification using RASTER data and self-organized neutral net-works. *International Journal of Applied Earth Observation and Geoinformation*, Elsevier, Vol. 7, No. 3, (2005), pp. 183-188.
7. Moreno M., Levachkine S., Torres M. and Quintero R.: Landform Classification in Raster Geo-Images, in A. Sanfeliu, J.F. Martínez-Trinidad, J.A. Carrasco-Ochoa (Eds.), *Lecture Notes in Computer Science*, Vol. 3287, Springer-Verlag, Berlin Heidelberg, (2004), pp. 558-565.
8. Nishii, R., Eguchi, S.: Supervised image classification by contextual AdaBoost based on posteriors in neighborhoods. *IEEE Transactions on GeoScience and Remote Sensing*, Vol. 43, No. 11, (2005), pp. 2547-2554.
9. Real Academia Española. 22ª edición. Madrid: Espasa Calpe, (2001).
10. Torres M., Guzman G., Quintero R., Moreno M. and Levachkine S.: Semantic Supervised Clustering to Land Classification in Geo-Images, in R. Khosla, R. J. Howlett, L. C. Jain (Eds), *Lecture Notes in Artificial Intelligence*, Vol. 3683, Springer-Verlag, Berlin Heidelberg, (2005), pp. 248-254
11. Torres M., Moreno M., Quintero R. and Fonseca F.: Ontology-driven description of spatial data for their semantic processing, in M.A. Rodríguez, I.F. Cruz, S. Levashkin and M. Egenhofer (Eds), *Lecture Notes in Computer Science*, Vol. 3799, Springer-Verlag, Berlin Heidelberg, (2005), pp. 242-249
12. Torres M., Moreno M., Quintero R., Guzmán G.: Applying Supervised Clustering to Land-sat MSS Images into GIS-Application, *Advances in: Artificial Intelligence, Computing Science and Computer Engineering, Research on Computing Science*, Vol. 10, (2004), pp. 167-176.
13. Wood, J.: The Geomorphological Characterisation of Digital Elevation Models, Ph. D. Thesis, Department of Geography, University of Leicester, UK, (1997).

Compressing Digital Elevation Models based on Run Length Encoding Approach

Giovanni Guzman and Rolando Quintero

Centre for Computer Research, National Polytechnic Institute, Mexico City, Mexico
{jguzmanl, quintero}@cic.ipn.mx
http://geo.cic.ipn.mx, http://www.cic.ipn.mx

Abstract. Up-to-date, some related algorithms to compress digital elevation models (DEM) or high-resolution DEMs use wavelet and JPEG-LS encoding approaches to generate compressed DEM files with good compression factor. However, to access the original data (elevations), it is necessary to apply a decompression approach to retrieve the contour lines. In this paper, we propose an algorithm oriented to compress a digital elevation model, which is based on a sequence of binary images to encode them using RLE compression technique, according to a specific height (contour lines). The sequence is compressed by applying a binary compressor. The main goal of our algorithm is that the specific parameters of the DEM (altitudes and contours lines) can be obtained without using a decompression stage, because the information is directly read from the compressed DEM. Our method reduces the amount of needed space to store DEM geo-images.

1 Introduction

The term digital elevation model or DEM is frequently used to refer to any digital representation of a topographic surface, however, most often it is used to refer specifically to a raster or regular grid of spot heights. Digital terrain model or DTM may actually be a more generic term for any digital representation of a topographic surface, but it is not so widely used [1]. DEM has gained popularity in applications to determine terrain's attributes, such as elevation at any point, slope and aspect, finding features on the terrain. The most important features are: drainage basins and watersheds, drainage networks and channels, peaks and pits and other landforms and, for modeling hydrologic functions, energy flux and forest fire [2].

On the other hand, the main purpose of a compression algorithm is to reduce the amount of information needed to describe the original data. To retrieve the original data we apply a process commonly named decompression. If the original information is fully-retrieved without any modification or alteration, then we can talk about a lossless compression algorithm. Also, it is possible to have compression algorithms with loss of information [9]. Many compression approaches first proceed to inspect the input image and try to detect different types of redundancy: statistical, psycho-visual or by correlation.

© M. Moreno-Ibarra, S. Levachkine (Eds.)
Semantics in Geoinformatics Applications
Research in Computing Science 25, 2006, pp. 73-84

A digital elevation model requires a huge amount of data, up to 100 or more megabytes of storage. Up-to-date, compression techniques are used to compress Digital Elevations Models with a high compression coefficient, in [4], [5] and [6] the authors propose to use wavelets and JPEG-LS encoding approaches. However, if the user requires to access to the file or to obtain certain information about the image, it is necessary to decompress all DEM information.

In this paper, we propose an algorithm oriented to compress a digital geo-image of a digital elevation model, which is based on a sequence of images with a specific height. The sequence is compressed by applying a binary compressor. The main goal of our algorithm is that the specific parameters of the DEM (altitudes and contours lines) can be obtained without using a decompression stage, because the information is directly read from the compressed DEM.

The rest of this paper is organized as follows: section 2 describes the proposed compression algorithm. Section 3 describes the manipulation of compressed DEM file to obtain both elevations and contours lines, without applying a decompression process. Section 4 presents the results obtained by applying our approach, and section 5 outlines the conclusions and future works.

2 Description of the compression algorithm

Nowadays, there are some types of DEMs such as: 7.5-minute, 15-minute, 2-arc-second, and 1-degree units [7]. For implementation purposes, we have proposed to choose the 1-degree DEM variant.

An important characteristic is that the frequency of the DEM is not high, because the pixels do not present changes in their structure; the value of each pixel is correlated with their neighbor values. Due to this, it is possible to apply this kind of compression. Moreover, this approach is not useful for common images.

Basically, a DEM file is integrated by three types of records, usually called A, B and C. The structure of these records is as follows [1]:

- *Record A* contains information, which defines the general characteristics of DEM, it includes descriptive header information related to the DEM's name, boundaries, units of measurement, minimum and maximum data values, number of type B records and projection parameters. There is only one type A record for each DEM file, and it appears as the first record in the data file.
- *Record B* contains elevation data and associated header information. All type B records of the DEM files are made up data from one-dimensional bands, called points. Therefore, the number of complete points covering the DEM area is the same as the number of type B records in the DEM.
- *Record C* contains statistics on the accuracy of the data in the file.

In a Digital Elevation Model, the altitude describes elevations greater or equal to 0 in a specific area. In consequence, the minimum altitude is 0 meters (sea level) and the maximum is 8,850 meters (Everest mount). In most of the cases, the DEM is displayed by using a 3-D render approach or applying a transformation function to obtain a gray level image. In Fig. 1 we depict an example of DEM file.

To detail the proposed algorithm, it is necessary to give some definitions that are important to describe our method.

The altitude[1] or height associated to one element that belongs to Record B in the DEM, is defined by Eqn. 1.

$$h(x,y)=\alpha, \tag{1}$$

To associate all points with same height α_k, we obtain the set defined by Eqn. 2.

$$S_{\alpha_k}=\{(x,y)\neg h(x,y)=k\}, \tag{2}$$

Fig. 1. Example of a DEM image

The number of elements in each S_{α_k} will be denoted by $card(S_{\alpha_k})$. Considering these definitions, the DEM altitudes are described by using Eqn. 3.

$$S=\{S_{\alpha_{\min}} \cup S_{\alpha_2} \cup ... \cup S_{\alpha_{\max}}\}, \tag{3}$$

where:

α_{MAX} denotes the maximum height in the original DEM.

Additionally, it is possible to define the join operation, according to Eqn. 4.

$$\bigcup_i^j S_\alpha=\{S_{\alpha_i} \cup S_{\alpha_{i+1}} \cup ... \cup S_{\alpha_{j-1}} \cup S_{\alpha_j}; i\le m\le j\}, \tag{4}$$

To obtain the difference between two sets, we can use the minus operation defined in Eqn. 5.

$$S_{\alpha_a}-S_{\alpha_b}=\{(x,y)\neg(x,y)\in S_{\alpha_a} \text{ and } (x,y)\notin S_{\alpha_b}\}, \tag{5}$$

With these definitions, we can formally state our proposed algorithm.

2.1 Compression approach

In this section, we provide a detailed explanation of the required steps to compress a 1-degree DEM with altitudes expressed as integer values.

Step 1. Records A and C are stored without modification in the output.

---

[1] We have adopted the convention that elevations in the DEM file are integer values.

Step 2. Find minimum and maximum altitudes $(\alpha_{\min}, \alpha_{\max})$ presented in the DEM and generate the header described in Table 1 in the output.

Table 1. Header description of the compressed DEM

Header of Compressed DEM	Type of variable to store information
Image width (columns)	integer (2 bytes)
Image height (rows)	integer (2 bytes)
$\alpha_{\min}$	integer (2 bytes)
$\alpha_{\max}$	integer (2 bytes)

The process starts with $k = \alpha_{\min+1}$ instead of $\alpha_{\min}$ due to all heights in the image are greater or equal to $\alpha_{\min}$, it is necessary to apply an iterative process from Steps 3 to 5.

Step 3. Determine the digital image associated to join operation, it is necessary to use the following two operations:

1. Compute the join operation composed of all altitudes lower than α_k, that is, inside the range $[0, k-1] \rightarrow i = 0, j = k-1$.
2. Generate a binary image applying the transformation function indicated by Eqn. 6.

$$g_{\alpha_k}(x,y) = \left\{ \begin{array}{l} 1 \text{ if } (x,y) \notin \bigcup_{i=\alpha_{\min}}^{k-1} S_{\alpha_i} \\ 0 \text{ otherwise} \end{array} \right\}, \tag{6}$$

In Fig. 2 we depict the result of digital images applying a threshold algorithm in Fig. 1, by means of the function described in Eqn. 6. The threshold applied values were $\alpha_k = 450$ and $\alpha_k = 575$.

(a) (b)

Fig. 2. Results of the threshold algorithm

3. If the new binary image is equal to previous altitude, it is necessary to skip the Step 4, and use the value -1 in the encoded DEM file and continue with Step 5.

Step 4. According to the binary image obtained in Step 3, we proceed to use the run length encoding (RLE) method [8] [10]. Basically, RLE is a straightforward way

of encoding data so that it takes less space. It relies on the string being encoded containing runs of the same character. In the RLE approach, it is important to indicate both, the frequency and the intensity encoded, but considering that the resulting image contains only two values (0 or 1), the value or intensity is not required in the final sequence. To determine the encoded image the following steps are applied.

1. By convention, the first frequency denotes the number of continuous zeros in the image.
2. By applying a top-down inspection[2], it is indispensable to count the number of pixels with same intensity and send this value to the encoded array, according to Eqn. 7.

$$value = \begin{Bmatrix} byte(value) \mid 80_{HEX} & \text{if } value < 128 \\ integer(value) & \text{if } value \geq 128 \end{Bmatrix}, \tag{7}$$

where:

byte (value) returns the byte representation of *value*, additionally we establish the most significative byte (MSB) to 1 using the Boolean OR operation.

integer (value) computes the integer representation of *value*[3].

By using this adjustment[4], it is possible to increase the compression factor because in some cases the total number of continuous pixels with same intensity should not be more than 128. To indicate this condition, the MSB of *value* is established to one; otherwise, an integer representation of *value* is stored in the new array.

3. Repeat the previous step until the end of the image is reached.
4. At the beginning of encoded array, we append the total number of encoded characters, which is the number of transitions between 0 and 1 in $g(x, y)$.

Step 5. Make a unitary increment to k variable, and repeat it from Step 1. The stop condition is used when $k > \alpha_{max}$.

In Figs. 3 and 4, we show the result of the compression approach applied to a small set of DEM elevations. Moreover, the original elevation data are presented in Fig. 3, and Fig. 4 depicts the final result.

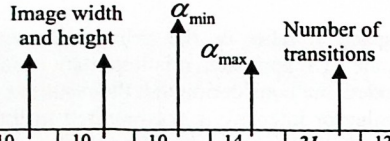
10	10	11	11	11	12	12	13	14	13
10	10	11	11	11	12	13	14	14	14
10	10	10	11	11	12	13	14	14	14
10	11	11	11	11	12	13	14	14	14
10	11	12	12	12	12	13	14	14	13
10	11	12	12	13	13	13	13	13	12
10	11	11	12	13	12	12	12	12	12
10	11	11	12	12	12	11	12	11	11
10	11	11	12	12	11	11	11	11	10
10	11	11	12	12	11	10	10	10	10

Fig. 3. Original elevations samples

[2] The inspection can be made in other directions, i.e. bottom-up (right-left and down-up).

[3] In this case, we consider that an integer requires 16 bits (2 bytes) of memory.

[4] For integers altitudes, the MSB always must be zero. Otherwise, it will not distinguish between bytes and integer elevations.



10	10	10	14	***21***	130	136	130	136	131
135	129	137	129	137	129	137	129	137	129
137	129	136	130	133	132	***23***	133	133	133
133	133	133	133	133	130	136	130	136	131
135	131	131	129	129	133	130	136	130	133
15	135	131	134	132	134	132	134	132	134
132	132	133	133	129	163	***11***	136	129	136
131	135	131	135	131	135	130	179		

Fig. 4. Compressed elevations

To represent information of the elevations, we require in the first case, 200 bytes of memory, and in the encoded version we only require 86 bytes.

2.2 Decompression method

The DEM compressed files contain three headers: Records A and C, DEM header, and the total of $\alpha_{max} - \alpha_{min} - 1$ RLE arrays. Retrieving original binary values from any RLE array do not provide the specific altitude at some point. If a point *p(x, y)* of source DEM has an altitude equals to *k*, this point has an intensity of 1 in total of $(k - \alpha_{min})g(x,y)$ images.

When we perform the join operation in the range $[\alpha_{min+1}, \alpha_{max}]$ a pixel *p(x, y)* has a 0 intensity if Eqn. 6 is not accomplished $([k+1, \alpha_{max}])$, then the gray-level in *g(x, y)* will be 1 $([\alpha_{min+1}, k])$. In conclusion the elevation of one point *p(x, y)* is determined by using Eqn. 8.

$$h(x,y) = \alpha_{min} + \sum_{i=\alpha_{min+1}}^{\alpha_{max}} g_{\alpha_i}(x,y), \tag{8}$$

In addition, the required steps of decompression stage are the following:

Step 1. Both Records A and C are retrieved directly from compressed file.

Step 2. Process the DEM compressed header, which contains the *width* and *height* of DEM image, as well as the minimum and maximum elevations. Generate an integer *width x height* matrix (referred as DEM image) and establish all the matrix values to minimum elevation.

Apply an iterative process from Steps 3 to 5 for each RLE encoded array.

Step. 3. Obtain the number of transitions in current encoded array (*n*), and remember that the value of *n* has been written before each RLE encoded array. If the value of *n* is equal to -1, then we apply the next steps with the first preceded valid RLE array.

Step 4. Both the absolute pixel position (*position* variable) and the current intensity encoded *(intensity)* are initialized to zero.

Step 5. Read the next byte (b_1) from DEM compressed file. If the most significative byte of this data is -1, then read next byte (b_2) and apply Eqn. 9 to obtain the value (n_0) of continuous encoded[5] intensities. In the compression algorithm, the first n_0 refers to zero intensity. Additionally,

$$n_0 = \begin{cases} b_2 + \left(b_1 \times (2\,\hat{}\,8)\right) \text{ if } MSB(b_1) = 1 \\ b_1 \,\&\, 7F_{HEX} \text{ otherwise} \end{cases}, \tag{9}$$

If the current intensity is equal to 0, we only increment with the value of n_0, the absolute position *(position=position + n_0)*. In other case, we need to make an n_0-unitary increment both the value of the DEM image at the position *p(x, y)*, and to the absolute position, as described in Eqn. 10.

$$\begin{aligned} x &= \left(position - width \times \lfloor position / width \rfloor\right) \\ y &= \lfloor position / width \rfloor \\ p(x,y) &= p(x,y) + n_0 \\ position &= position + n_0 \end{aligned}, \tag{10}$$

Change the value of intensity if the current value is 0, set it to 1, and to 1 else, repeat Step 5 until process all RLE values.

3 Compressed DEM management

In section 1, we cited the most important characteristics of this compression algorithm. We assume that it is not indispensable to apply a decompression method to obtain information about the elevation from the DEM (read data). Additionally, we can generate the contour layer, which has several map applications related to Geoprocessing area. These two operations are described in the next section.

3.1 Accessing data elevations

In some cases a DEM file requires up to 100 MB storage space. Moreover, some person can require an application that works with 10, 20, or more DEMs. It is possible that any user does not have storage limitations, but in conditions that the use of the lower memory, or the use of the hard disk space is crucial, the DEM compressed file could be useful. To retrieve the elevation at some point *p(x, y)* we cited Eqn. 8, however it requires to process all RLE encoded array. To reduce processing time, we can apply the Eqn. 11.

$$p(x,y) = \alpha_h \neg g_{\alpha_h}(x,y) \neq 0, \left(g_{\alpha_{h+1}}(x,y) = 0 \vee \alpha_h = \alpha_{\max}\right), \tag{11}$$

[5] We suppose that high-part of the integer is stored first.

This equation basically consists of a lineal-search in all RLE encoded arrays, in the cases when the desired elevation is near α_{min+1}, the elevation is rapidly obtained, otherwise all arrays (closer to α_{max}) will be processed. In consequence, we have $\Theta(n/2)$ algorithm, where n denotes the total number of bytes to process. To improve this process the *bin-search* can be used. When *bin-search* starts, the total RLE arrays will be equal to n (see Eqn. 12).

$$A = \{a_0, a_1, ..., a_n\}, \tag{12}$$

Formally, the binary-search function (BS) is described in Eqn. 13.

$$BS(A, k_1, k_2, m, g_{a_m}(x,y)) = \begin{cases} \alpha_{\min} + k_1 + 1, \\ \text{if } g_{a_{k_1}}(x,y) = 0, g_{a_m}(x,y) = 1, k_2 = m+1 \\ \\ \alpha_{\max} \text{ if} g_{a_{k_2}}(x,y) = 1, k_2 = n \\ \\ BS\left(\{a_{k_1} \ldots a_{m'}\}, k_1, m'-1, m', g_{a_{m'}}(x,y)\right), \\ \text{if } g_{a_{m'}}(x,y) = 0 \\ \\ BS\left(\{a_{m'+1} \ldots a_{k_2}\}, m'+1, k_2, m', g_{a_{m'}}(x,y)\right), \\ \text{if } g_{a_{m'}}(x,y) = 1 \end{cases}, \tag{13}$$

where:

A denotes the RLE arrays to search.

k_1 and k_2 are the number of RLE array $k_1, k_2 \in \{0, 1, \ldots, n\}$.

m is the middle value of previous binary-search.

$g_{a_m}(x,y)$ is the altitude at middle position in previous search.

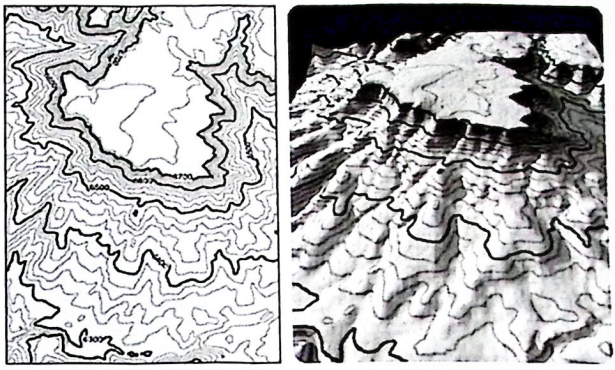
m' is the new middle value, computed as $m' = \lfloor (k_2 - k_1)/2 \rfloor + k_1$.

In the first execution of the search function $k_1 = 0, k_2 = n, m = 0, g_{a_m}(x,y) = 0$. A special condition is reached when the k-RLE array is empty because $g_{a_k}(x,y) = g_{a_{k-1}}(x,y)$, to solve this ambiguity the method searches for the first previous RLE array, which is not empty.

3.2 Generating contours without compression

Contours are lines drawn on a map, which connect points of equal elevation. Contour lines are useful because they allow us to show the shape of the land surface (topography) on a map [3]. The maps depicted in Fig. 5 illustrate an example of

contours. Fig. 5 (a) shows the contour lines for some region and in Fig. 5 (b), a 3-D perspective of same area overlapped with contours is depicted.



(a) (b)

Fig. 5. Contour lines example: (a) Only contours data (b) Contour lines overlapped with 3-D model

The methodology to obtain contour lines form in a Digital Elevation Model is not complicated; some commercial tools to handle cartographic data make this task. However, when we work with DEMs, these systems require all elevation date, but our proposed algorithm generates the contour lines directly form compressed DEM file. In this section we point out this process.

The single required parameter is the altitude interval to sample DEM (Δ), all additional information is available form compressed data.

Step 1. Generate an $N \times M$ matrix and establish all values to 0, this matrix will represent the contour lines image $f(x, y)$. The size of the matrix is obtained from DEM compressed header.

Step 2. Take the first RLE encoded array (i.e. encoded data of $g_{\alpha_{\min+1}}$). Obtain original binary image by using Eqn. 8. If the array is empty, then apply the same criteria as we defined it in previous sections. With original binary image, it is necessary to obtain its negated version.

Step 3. Compute the 8-connected contour of binary image. Strictly, an object pixel $p(x, y)$ is part of contour, if the number of 8-neighbors with intensity equal to zero (background pixels) is at least equal to 1.

Step 4. Remove redundant contour pixels. All pixels that accomplish one of the masks defined in Fig. 6 are removed.

X	1	X
1	1	X
X	X	0

X	1	X
X	1	1
0	X	X

0	X	X
X	1	1
X	1	X

X	X	0
1	1	X
X	1	X

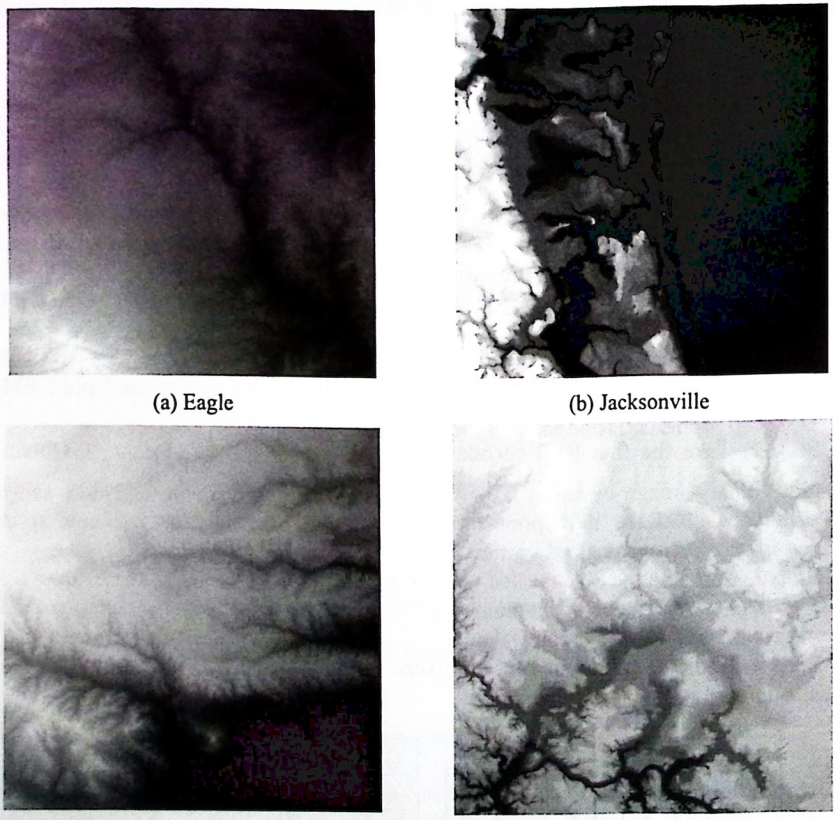
Fig. 6. Set of redundant pixels in detection masks

Step 5. Establish value of one at coordinates *(x, y)* in contour lines image, if the pixel *p(x, y)* is a member of non-redundant contour set obtained in the Step 4.

Step 6. Process each remained encoded array by applying from the Step 4.

4 Tests and Results

In this section, we show in a detailed way, the obtained results with our proposed approach, and compare it with commercial WinZip program (it uses LZW coding method); in this case the standard-compression option has been applied. The DEMs used for the tests are displayed in Fig. 7, and all DEM sources have the same spatial resolution (1,200 x 1,200 pixels) and their sizes are 10MB approximately. In Table 2 the numerical information about the obtained results appears.



(a) Eagle (b) Jacksonville

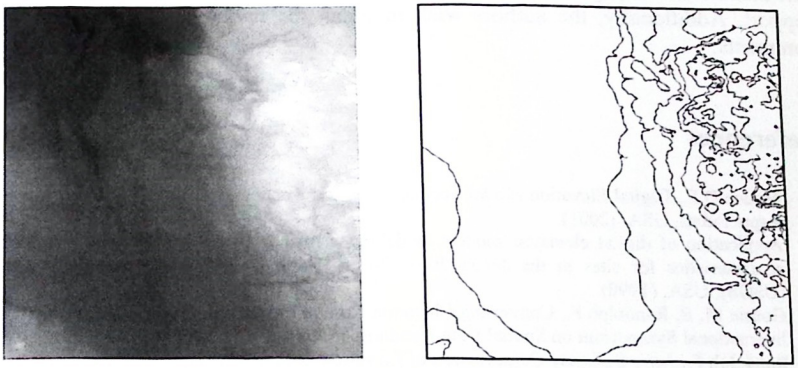
(c) Pratt (d) Cincinatti

Fig. 7. DEMs samples used in compression tests

Table 2. DEM compression results[6]

Name	WinZip	Compression	Our Technique	Compression
Eagle	917,504	90.67%	1,645,939	83.27%
Jacksonville	262,144	97.33%	236,126	90.07 %
Pratt	868,352	91.17%	1,306,272	86.73 %
Cincinatti	819,200	91.67%	1,642,829	83.31 %

Regarding the contour line computation, in Fig. 8 (b) we show the obtained result with the DEM source that appears in Fig. 8 (a). The altitude interval is 25 meters.



(a) (b)

Fig. 8. Contour lines result: (a) DEM original geo-image (b) Contour lines obtained from DEM compressed file

5 Conclusions

In the present work, we develop a loss-less compression algorithm using RLE encoding approach. Although the compression factor is not the same, we can obtain this factor applying other techniques, it is important to mention that encoded elevations can be directly read from DEM compressed file, and the computation of contours lines is easy and fast. The most important part of the test is the compression factor, which has been 80% high. To make some enhancements, this algorithm should be adapted as new format to store and describe digital elevations models. Future works are oriented to study others, encoding formulations that allow us to increase the compression coefficient and attempt to implement developed compression algorithms to compare them with our method.

An important characteristic is that the frequency of the DEM is not high, because the pixels do not present changes in their structure; the value of each pixel is

[6] All data are expressed in bytes.

correlated with their neighbor values. Due to this, it is possible to apply this kind of compression. Moreover, this approach is not useful for common images.

On the other hand, we have considered the most important characteristics of digital elevation models in our method, because it is important to point out that DEMs are complex geo-images, which involve several properties of a certain environment. In addition these properties reflect the semantics of the digital elevation models.

Acknowledgments

The authors of this paper wish to thank the CIC, CGPI, CONACYT and IPN for their support. Additionally, the authors wish to thank the reviewers for their pertinent comments.

References

1. Maune D.F., Digital Elevation Model Technologies and Applications: The DEM Users Manual, Asprs Pubns, USA, (2001).
2. Application of digital elevation models to delineate drainage areas and compute hydrologic characteristics for sites in the James River Basin, North Dakota, U.S. Geological Survey (USGS), USA, (1990).
3. Gousie M. B, Randolph F., Converting Elevation Contours to a Grid, Proceedings of the Eighth International Symposium on Spatial Data Handling, (1998).
4. Randolph F., Amir S., Lossy Compression of Elevation Data, USA, (1995).
5. Shantanu D. R., Sapiro G., Evaluation of JPEG-LS, the New Lossless and Controlled-Lossy Still Image Compression Standard, for Compression of High-Resolution Elevation Data, *IEEE Transactions On Geoscience And Remote Sensing*, Vol. 39, No. 10, (2001), pp. 2298-2306.
6. Creusere C. D., Compression Of Digital Elevation Maps Using Nonlinear Wavelets, IEEE, (2001), pp. 824-827.
7. Standards for Digital Elevation Models – Part 1, U.S. Geological Survey (USGS), USA, (2002).
8. Hinds, S.C, et al, A document skew detection method using run-length encoding and the Hough transform, IEEE, Proceedings, 10th International Conference on Pattern Recognition, Vol. 1, (1992), pp. 464-468.
9. Gonzalez R, Digital Image Processing Second Edition, Prentice Hall, USA, (2001).
10. Beenker, G F M, Immink, K A S, Generalized method for encoding and decoding run-length-limited binary sequences, *IEEE Trans. Info. Theory*, Vol. IT-29, No. 5, (1983), pp. 751-753.

Geospatial Analysis and Its Applications

Applying Data Mining to Determine Relationships between Relief and Precipitations: A Case of Study In Northern Chile

Marco Seron[1], Pamela Soto[1], Claudio Meneses[1],
Rodrigo Riquelme[2], and Jorge Quezada[2]
Universidad Católica del Norte, Antofagasta, Chile

[1] Department of Systems & Computer Engineering
{maseron, pamelasoto, cmeneses}@ucn.cl
[2] Department of Geologic Sciences
{rriquelme, jquezad}@ucn.cl

Abstract. In this work we are interested in identifying aspects of the relief that can be associated to tectonic or climatic factors. Applying Data Mining techniques on a group of attributes generated from the transformation of longitudinal profiles of a topography using Wavelet Transform and Matching Pursuit and the historical data of the average precipitations of the last 30 years. The study considers the region among the latitudes 18°S and 27°S in the coastal zone of northern Chile. The objective this study is to find any interesting relationship among latitudinal variations of the characteristics of the relief of the Mountain range of the Coast and the latitudinal climatic variations in northern Chile. The results obtained and an analysis of them indicates no evidence of lineal correlation among relief and climatic variations on this zone, although association rules show certain relationship between relief and precipitations, but with low confidence.

1 Introduction

The role of climate a tectonic on the landscape evolution is widely documented in geomorphology. Tectonic together with climate are the most important factors determining the rate at which a landscape is eroded. In this context, several investigations analyze the landscape topography in order to determinate the relative impacts of these factors in the landscape configuration (e.g. [20], [6], [17]).

Nowadays, availability of Digital Elevation Model (DEM) allows us to explore and to extract some quantitative characteristics of the topography, which can be interpreted in terms of tectonic and climatic factors operating on the landscape [3]. However, the establishment of the impact of climate (or/and tectonic) on the landscape frequently remain uncertain because no formal rules relating landscape characteristics and climate (or tectonic) are used.

In this paper we propose and apply an objective and effective methodology concerning to the discrimination of the relative impact of the climate and tectonic on the landscape configuration. We choice the Coastal Cordillera of the Atacama Desert

(northern Chile), a mountain range extending more than 1000 km between the latitudes 18°S and 28°S, in the hyper-arid northern Chile. Three factors makes of the Costal Cordillera an excellent place to carry-out our objectives: (1) its topography exposes accurate NS oriented segment which can be, a priori, correlated to the segment of the Atacama Fault System (AFS), a discontinuous set of faults striking parallel to the Coastal Cordillera [22]. (2) It is effected by overall uplift tectonic conditions during the Quaternary. (3) Hyper-arid conditions prevail at least from late Pliocene [13] and, a perceptible NS mean annual rainfall gradient define the present-day climatic conditions of Coastal Cordillera [8].

We address to individualize which of the landscape characteristics of the Coastal Cordillera are related to tectonics factors and which of them can be associated to climatic conditions. To analyze the Coastal Cordillera landscape we consider NS oriented topographic profiles defining mathematical signals, where amplitude corresponds to topographic altitude varying again latitude (from north to south). In this signal, low frequency variations represents variations in large-scale landscape characteristics (e.g. segments of the landscape with different mean altitude). Whereas, high frequency variations represent latitudinal variation of incision, i.e. north-south successions of valleys and interfluves. We suspects that low frequency variations can be related to differential latitudinal landscape tectonic control associated to the different segment of the AFS. Whereas, high frequency variations evidence a greater capacity to incise the landscape that responds to the NS mean annual rainfall gradient or latitudinal variations of the uplift rates: a latitudinal increase of mean annual rainfall or uplift rate inducing an increase of amplitude and frequency of incision.

In order to test our hypotheses, we analyze the topographic signals by considering data transformations performed over a derived group of attributes. These transformations are carried out using the Wavelet Transform (WT) and Matching Pursuit (MP) techniques.

With the topography one can obtain information on the position of an element with regard to the longitude, latitude, and its elevations, among others. The way in that the precipitation erosions a landscape responds to a mathematical model [17]. Based on the mentioned above, we can think that observing the relief, we may determine the magnitude of mean annual rainfall that affected it.

From the Data Mining perspective, one of opposing problems when working with satellite images, is to find some transformation for these images. This transformation should be the more appropriate to extract the information associated to the problem.

For this problem, mainly two sources of information are available: satellite images and data of average coastal precipitations of the last 30 years in northern Chile. With this information, we want to explore the existence of significant relationships between the relief and the precipitations in the region limited between the latitudes 18,5°S (Arica) – 27°S (Caldera), and between the Coast Mountain and the beginning of the Intermediate Depression.

In addition to the main goal of identifying interesting relationships among relief and climate or tectonic factors, we consider the following specific goals:

- To confirm or to discard the hypothesis that the distribution of those systems of drainage of the Coast Mountain are related with the variations of latitudinal precipitations.

- To find any interesting correlations between topography and precipitations in the zone limited from Arica to Caldera and from Coast Mountain to Intermediate Depression of northern Chile.
- To carry out an spectral analysis of the topography, in order to verify whether the topography determines information on precipitations.

2 Methodology

In order to replicate this study to other zones, we formulate a guide (sequence of phases and tasks), based on the methodology propose to carry out a Data Mining project [10]. Our guide (see Fig. 1) is composed of the following phases:

- **Data Preparation:** In this phase, we carry out all necessary data preparation tasks over the two sources of data: satellite topography data and precipitations records corresponding to the target zone. In this phase, the followings tasks should be performed:
  - *Obtaining Longitudinal Profiles:* From the topography data, longitudinal profiles are traced along the Coast Mountain. These profiles must have a constant distance among them, and also the first profile must have a constant distance to the coastal border. This operation is performed by using a GIS tool, storing the results in a data file.
  - *Error Correction:* After the profiles are traced, they may pass on topography defects. Therefore, a process of exploration and correction of these defects should be performed if any defect is detected.
  - *Identification of Structural and Lithological Elements:* It consists of the identification of structural and lithological elements of diverse sizes, belonging to the region under study, with the purpose of establishing their relationships with the results to be obtained.
  - *Interpolation of Precipitations:* Starting with the data points for which precipitations are already known, we should interpolate any required intermediate point between two known points, along the zone under study. This with the purpose of having continuous climatic data that can be subsequently correlated with topography data. Any suitable interpolation method should be considered (e.g., lineal interpolation, Lagrange interpolation).

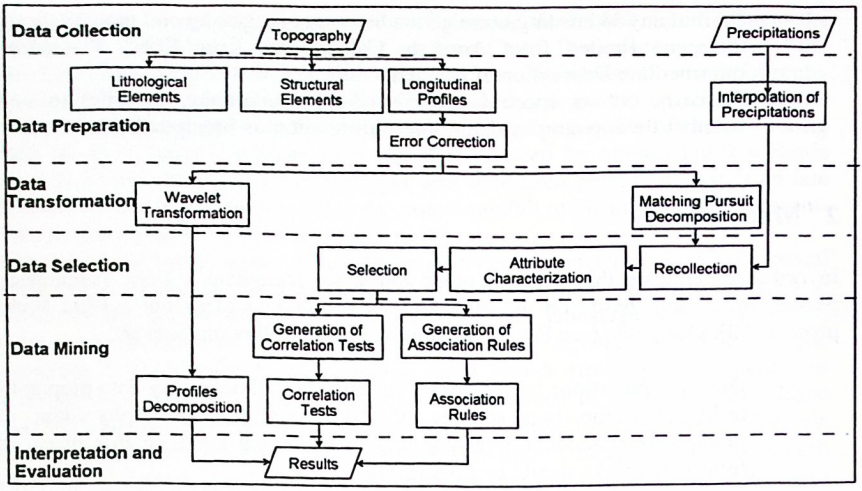


Fig. 1. A schematic overview of the methodology

- **Data Transformation:** Once the profiles are generated, functional transformations are applied in order to decompose them. These transformations map profile data into signal data. In our case, we use the following two transformations:
  - *Wavelet Transform:* The purpose is to obtain a visual decomposition of the profiles.
  - *Matching Pursuit* Decomposition: The purpose is to obtain numeric values of the frequency components represented by the longitudinal profiles.
- **Data Selection:** Once the data have been transformed, it is necessary to identify which one will be finally used in the following phases of the process. Activities that are carried out in this phase are:
  - *Recollection:* The data generated in the transformation and interpolation tasks must be integrated, with the purpose of having a set of possible attributes and records.
  - *Attribute Characterization:* All attributes should be characterized in terms of basic properties such as distribution of values, data quality, statistical information, etc.
  - *Selection:* Once the data are characterized, the relevancy of each attribute should be estimated, using either expert assessment or algorithmic approaches (e.g., filter and wrapper attribute selection methods), or both. In any case, attribute relevancy may depend on the region under study, given that some attributes may be more o less important.
- **Data Mining:** Based on the prepared, transformed, and selected data, data mining techniques are applied in order to generate several candidate models.

In our case, we applied two suitable techniques for the problem of finding relationships in the data:

- *Generation of Correlation Tests:* Statistical tools are applied in order to quantify possible correlations between precipitation data and profile decomposition data, this with the purpose of identifying any dependency between latitudinal variation of the precipitations and the relief of the zone under study.
- *Generation of Association Rules:* Data mining tools are applied in order to generate rules that show significant associations among attribute values. This algorithmic approach may discover no lineal dependency between latitudinal variation of the precipitations and the relief.

– **Evaluation and Interpretation of Results:** The final phase is to evaluate the results (e.g., lineal correlation models, association rules, Wavelet Transform profiles) both individually and crossing the information among them. In addition, it is very important that the results be interpreted by the domain expert and the user in order to have a qualitative estimation whether the results are or not interesting from the expert point of view.

3 An Overview of Applied Techniques

In this section, we overview the main techniques used during the development of the case of study, with the purpose to make the paper as self-content as possible. We describe Matching Pursuit and Wavelet Transform techniques used to decompose profiles (signals), as well as lineal interpolation, lineal correlation, and association rules.

3.1 Matching Pursuit

Matching Pursuit process consists in finding a function f as the pondered sum of elements obtained from a Gabor dictionary [18]. In others words, f is represented as:

$$f = \sum_{\gamma \in \Gamma} \alpha_\gamma g_\gamma \tag{1}$$

This process is used to carry out a decomposition of one signal in different components of frequency that added recompose the original signal. The comparison of the components of frequencies in the Matching Pursuit process is carried out using the Short Time Fourier Transform (STFT) [12], also known as Gabor Transform (GT). STFT cuts the signal in small pieces, which are known as windows. The form of the resulting components of decomposition is strongly bound to the selected window type, which is associated to the fact that the reconstructed signal can have bigger or smaller error.

To decompose this signal, we used the Guimauve software tool [2], which in turn uses the technique of Matching Pursuit to generate the decomposition of the signal

into a group of signals located in different time-frequency ranges. Guimauve is based on a group of software libraries, whose name is LastWave [1].

The Guimauve software tool allows to decompose the signal in the following elements:

- **Atoms,** which indicate in how many signals the decomposition of the original signal was carried out. For example, if the quantity of atoms is 200, this indicates that the signal was decomposed in 200 signals.
- **Coefficient,** which shows the coefficient of decomposition of the signal using STFT.
- **Octave,** which is an interval of frequencies in a power of two.
- **Time,** which indicates in that time a certain frequency is found.
- **Frequency,** which indicates the frequency of the decomposed signal.

3.2 Wavelet Transform

The Wavelet Transform [16] carries out a filtrate of the original signal, using filters that allow us to pass a certain range of frequencies either below a threshold (low filters) or above that threshold (high filters) [2], and that eliminate certain components of frequency of the signal. This procedure is repeated with the signals generated from the previous filtrate. The process is carried out mainly with two approaches: The Continuous Wavelet Transform (CWT) and the Discrete Wavelet Transform (DWT).

- **CWT:** It uses a window that frames the signal inside an interval, concentring the analysis only in a segment of the signal. It tries to express a continuous signal X(t) in the time, by means of an expansion of proportional coefficients to the internal product between the signal and different scaled and transferred versions of a prototype function $\Psi(t)$ (mother wavelet) [9]. The CWT is defined as:

$$CWR\ (a,b) = \frac{1}{\sqrt{a}} \int X(t) \Psi \frac{(t-b)}{to} dt \tag{2}$$

  where: a controls the width of the function (scale variable) $\Phi(t)$ and b gives us the location in the domain of the time (translation variable). The scale variable allows us to dilate or to compress the mother function to control the resolution that is wanted to obtain in the transformation. The translation variable controls the location of the function $\Psi(t)$ in the time space. This allows us to displace the mother function of scale a through the signal. With this one can say that to smaller size of a (small scale), a bigger quantity of translations must be performed, in order to cover the complete signal.
- **DWT:** This transformed part of the base that the scale parameters and translation are discrete, this with the purpose of simplifying the quantity of calculations to carry out. Under this assumption, the following thing is generated:

$$\Psi_{a,b}(t) = \frac{1}{\sqrt{a}} \Psi \frac{(t-b)}{to} to, b \in \Re, a \neq 0 \tag{3}$$

with $a=2^{-j}$ and $b=k2^{-j}$, it becomes the group of functions:

$$\Psi_{j,k}(t) = 2^{\frac{j}{2}} \Psi(2^{j} - k) j, k \in Z \quad (4)$$

This decomposition generates a set of elements, d12 to d1, from small to big frequencies, and a12 corresponds to the first decomposition (also called base decomposition).

Once we know the definitions of both ways of transforms, it is necessary to define the mother function more appropriate for the problem.

The advantage of Wavelet over STFT for this study is that STFT lacks of properties (orthogonality, conservation of the energy, entropy) necessary to be able to make a correct interpretation of the decomposition. To decompose the signal by means of Wavelets, we used Wavelet Toolbox of Matlab 7.0. On the other hand, Wavelet and Matching Pursuit are complementary techniques for this study. The first one allows us to visualize constituent frequencies of the train of signals that can be quantified through the second one. The visual results obtained with Wavelets are corroborated by the quantitative results obtained with Matching Pursuit.

3.3 Correlations

Correlations are used to examine the degree of similarity of the values of two numeric variables. A standard formula to measure lineal correlation is the correlation coefficient r, which is a numeric real value between -1.0 and 1.0. If r is positive then the variables are correlated positively, i.e., the variables have a similar behavior (both growing or falling at the same time); if r is negative then the variables are correlated negatively, i.e., when a variable grows the other one falls [14]. On the other hand, when r is 0, there is not correlation.

3.4 Association Rules

They express behavior patterns among the data in function of the combined appearance of values of two or more attributes. The main characteristic of these rules is that they try with nominal attributes, to difference the correlations that make it with the numeric ones. In short, these rules express combinations of values of the attributes that happen more frequently. They can be expressed in many ways, although the most common form is of the style: "if the attribute X takes the value d then the attribute Y takes the value b", which usually is written as: $X=d \rightarrow Y=d$. An association rule is usually worked with two measures to know the quality of the rule [14]:

- **Support:** It is defined as the number of instances that the rule predicts correctly.
- **Confidence:** Also denominated precision, it measures the percentage of times that the rule is met when it can be applied.

3.5 Interpolation

- **Lagrange Interpolation:** Its is defined of the following form:

$$p(x) = y_0 l_0(x) + y_1 l_1(x) + \ldots + y_n l_n(x) = \sum_{k=0}^{n} y_k l_k(x) \quad (5)$$

where l_0, l_1, ..., l_n are polynomials that depend only on the sampled values x_0, x_1, ..., x_n, but they do not depend on the values y_0, y_1, ..., y_n. The general formula for the polynomial li is:

$$l_i(x) = \prod_{j=0, j\neq i}^{n} \frac{x - x_j}{x_i - x_j} \quad (6)$$

Using these polynomials in the equation (5) we obtain the exact form of the Lagrange interpolation polynomial [4].

- **Lineal Interpolation:** It is the adjustment of one straight line between two known points, to determine the value of an unknown point that is among both previous ones. Be F(x) a function for we know their values $f(X_0)$ and $f(X_1)$ in the points X_0 and X_1. The more simple approach to approximate the f(x) values for any x belonging to the interval $[X_0, X_1]$, is to consider the polynomial of first degree [4]:

$$F(x) = f(x_0) + \frac{x_1 - x_0}{x_1 - x_0}(x - x_0) \quad (7)$$

4 Sources of Information

4.1 DEM

A DEM [19] is a representation of the landscape of the form **(x, y, z)**, which correspond to the latitude, longitude, and elevation, respectively. The precision of these data depends on two factors: the horizontal precision and the vertical precision.

- **Horizontal Precision.** It corresponds to the resolution of the satellite image from which the topography was obtained. In the case of the DEM used for this study, it has a resolution of 90 meters. That means that each pixel of the image represents a square of 90 squared meters.
- **Vertical Precision.** It corresponds to the precision calculated from data obtained in landscape of elevations, which are compared with the data of the elevations obtained from the DEM. This operation is carried out using the Medium Squared Error (MSE), with at least 30 points, 20 of which correspond at the interior of the DEM, and 10 to those correspond to the border of the same one.

$$MSE = \sqrt{\frac{\sum (Z_i - Z_i)^2}{n}} \quad (8)$$

From the DEM data, we carry out longitudinal profiles in four places, in order to appreciate the variation of the relief when we advance from the coast toward the intermediate depression. These cuts are shown in Fig. 2A, where the first, second, third and quarter profile are shown from left to right. With the information obtained from the profile, we can build scatterplot graphics among the latitude in a point (axis X) and the elevation of the same one (axis Y), resulting in the dispersion graphics shown in Fig. 2B.

These graphics show the form of the relief through the cuts, and where depressions, plains and elevations are appreciated. These profiles can be considered as signals. This since the elements that compose it can be transformed to homologous elements, but in the domain of the signals, and where the elevation represents the width, the latitude can be homologated with the time, and the variations in the elevations of the relief indicate different frequencies in this signal, to which we will call topography.

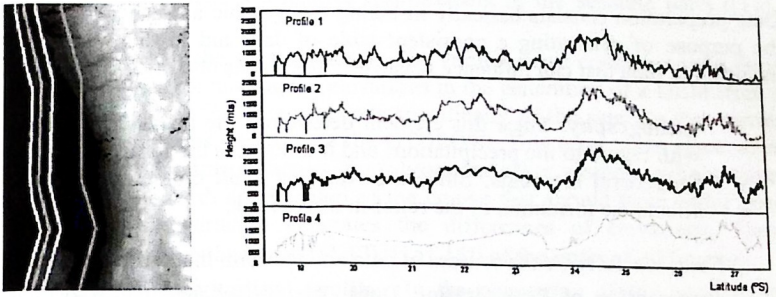


Fig. 2: A) Longitudinal profiles used in the study, B) Scatterplots of the longitudinal profiles.

4.2 Precipitations

The data associated to precipitation correspond to coastal cities located among latitudes 18,5°S and 27°S (see Table 1). These data of precipitations are an average of the last thirty years [8]. Given that the necessary data should be from the latitude 18,5°S until the latitude 27°S, we carry out an operation of interpolation of precipitations in all range of latitudes, with the purpose of being able to use these data to establish correlations and associations.

Over this dataset of precipitations, a lineal interpolation (7) is performed among the diferent data points. In this case, the Lagrange interpolation generates points with smaller values than those in the Table 1. As shown in Fig. 3 (a), the curve generated by the Lagrange interpolation gives negative values between 20°S and 22°S, and some points (e.g., Tocopilla and Chañaral) are quite different from the true ones. On the other hand, as shown in Fig. 3 (b), lineal interpolation does not generate negative

values, and it fits the initial values. For this reason, the interpolation of precipitation values is performed by using the lineal method.

Table 1. Precipitations in different cities of the coast of northern Chile.

City	Latitud(°S)	Precipitation (mm)
Arica	18,5	0,5
Iquique	20,8	0,6
Tocopilla	22	2
Antofagasta	23,5	1,7
Taltal	24,4	8
Chañaral	26,2	12
Caldera	27	27
		Mean: 4,13 Standard Deviation: 4,75

5 Data Preparation

Data preparation consists basically in taking topographic and precipitation data with the purpose of generating a consistent table of data and in this way to avoid the inclusion of data that can influence negatively the results obtained.

- **Lithography**. Since this element determines the behavior of the landscape with regard to the precipitations and it influences the form that is modeled.
- **Structural elements**. Since the Atacama Fault crosses the profiles and it affects the formation of the relief in direct form.

Next, we describe the sources of data and the form that these are preparated:

1. **Preparation of Precipitation Data**. We applied lineal interpolation to the precipitation data to approximate unknown values. As we can observe, the lineal interpolation is well adjusted to the group of precipitation data (see Fig. 3 and Table 1).

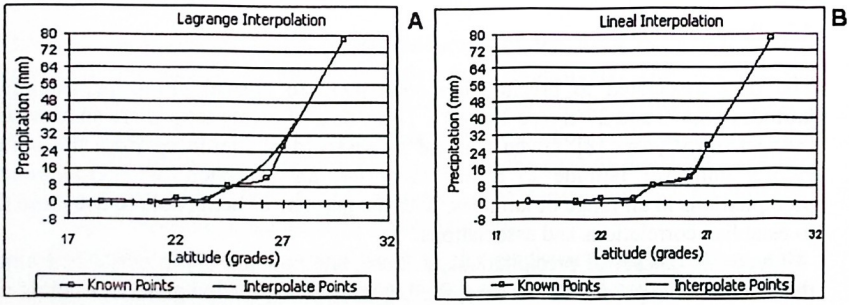


Fig. 3: (A) Scatterplot of the lagrange interpolation of the precipitation data. (B) Scatterplot of the lineal interpolation of the precipitation data.

2. **Determination of lithological elements.** As previously mentioned, the signal "Topography" is made up of different elements that can be reflected in different components of frequencies. Among these elements are rocks of diverse sizes. Also, maps were generated to identify bodies by using convolution filters. Specifically, we used the filter of Means to highlight the bodies of a certain size. The filter carries out a substitution of the value of a pixel for the mean of their k-nearest neighbors, and where the size of the bodies that we want to identify determines the value that k should take. After obtaining the result of the filter application, a subtraction is made among the resultant map and the original one. This operation generates a map that highlights the bodies of certain size. These outstanding bodies are associated to certain type of rocks in the lithography of northern of Chile.
3. **Determination of structural elements.** In the same form as some bodies of different sizes was identified, it is necessary to isolate the components of frequency of the associated signal "Topography" to elements of structural control, specifically the faults, which are discontinuities formed in the rocks for fractures when concentrations of tectonic forces exceed the resistance of the rocks, that generates a relative displacement of the separate parts ([11], [26]).
4. **Determination of the vertical error.** The vertical error consists on the error introduced when calculating the elevations in the generation of a DEM from a satellite sampling. This error depends specifically of the satellite that carries out the mensurations, in our case the GTOPO90. This error takes place when for atmospheric, climatic factors, floor types and other reasons, radio's waves sent toward the earth are deviated of the course that should have under good conditions, this variation generates the differences of elevations. These vertical errors are estimated ([25], [7], [15], [5]) of the order of ±25 meters.
5. **Correction of longitudinal profiles.** In the generation of the profiles, we detected the presence of sectors in which the elevations were zero, this corresponds to errors in the generation of the base satellite image. This was detected in the exploration of the data of each profile. To overcome this problem, we calculated the mean value between the borders of the hole, and we put this value in the corresponding position of the missing values. Afterwards, the same operation is carried out between the border and the inserted value. In this way, we fill in missing values of the generation of the profile.

6 Data Transformation

The decomposition of the signal "Topography" seeks to identify which of the obtained components of frequency is related with the latitudinal precipitations in the region mentioned before. The signal decomposition process was carried out by means of the use of Matching Pursuit and Wavelets methods, with the purpose of comparing the results of both methods. We used Matching Pursuit for the evaluation of the association rules and correlations, by means of the Guimauve tool. When STFT is

used it should be indicated which is the type of sampling window necessary to generate the dictionaries of signals.

We carried out a group of decompositions of the signal corresponding to profiles 1, 3 and 4, in way of seeing which are the errors generated to the moment of reconstructing this signal. In Fig. 4 we can see the reconstruction error for the first profile. The objective of this test is to determine which is the more appropiate window for the decomposition of the signals

The behavior of the error shown in Fig. 4 is similar in all the analyzed profiles. Also, we can observe that the Blackman window type is the one with worse results, and that the Hanning and Spline 1 window types show the best results when trying to reconstruct the original signal.

Finally we use the Spline0 type, that corresponds to the family of the B-Spline and is known as rectangular ([24], [23]). The rectangular form of this window allows us to appreciate clearly signal components along the time, without deformations, which facilitates the task of interpretation of each component.

We used Haar Wavelets to identify in to graphic way the "blocks" of frequencies. This Wavelet mother facilitates the visual comparison of the decomposition with those "blocks" of frequencies of the topography and it also allows to compare it with the decomposition accomplished by means of rectangular Spline in Matching Pursuit. The tool used to generate this decomposition was Wavelet Toolbox of Matlab 7.0.

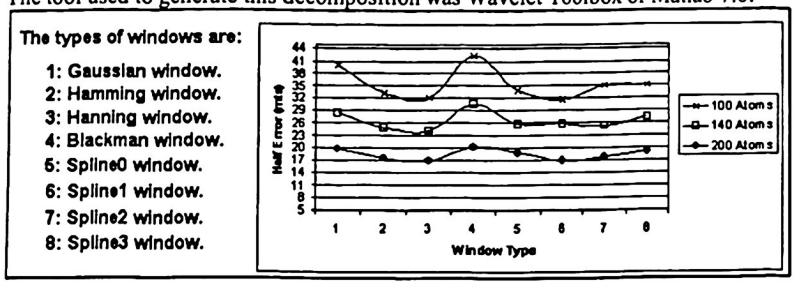


Fig. 4: Error for the different windows.

7 Data Selection

7.1 Compilation of Attributes

As a result of all the mentioned above, we obtain the following attributes:

- **Longitude (L).** It corresponds to the longitudinal location of an area, inside the study area.
- **Initial Latitude (LI).** It corresponds to the initial latitudinal location of a component of frequency.

- **Final Latitude (LF).** It corresponds to the final latitudinal location of a component of frequency.
- **Latitudinal Distance (LD).** It corresponds to the difference between Initial and Final Latitude, and it is measured in degrees.
- **Frequency (F).** It corresponds to the value of a component of the signal.
- **Precipitation (P).** It corresponds to the quantity of precipitations in an area of the study region. They are measured in millimeters of water fall.
- **Elevation (E).** It corresponds to the values of the elevations in the study region, measured in meters.

7.2 Characterization of Attributes

The candidate attributes are characterized of the following form (see Fig. 5):

A

Name	Type	Units	Possible Values
Longitude	Numeric	Degrees	Min.: 69,7767 Max.: 70,8482
Initial Latitude	Numeric	Degrees	Min.: 18,3351 Max.: 27,5221
Final Latitude	Numeric	Degrees	Min.: 18,4227 Max.: 27,5278
Latitudinal Distance	Numeric	Degrees	Min.: 0,0024 Max.: 5,8619
Frequency	Numeric	No unit	Min.: 0 Max.: 0,25
Precipitation	Numeric	Millimeters	Min.: 0,503868 Max.: 36,23628
Elevation	Numeric	Meters	Min.: 40 Max.: 2650

B

Attribute	Mean	Standard Deviation
Latitude	23,1889508	2,714656427
Precipitation	8,402134897	10,74030827
Frequency	0,011599665	0,011238048
Elevation	959,5321571	482,6221739

Fig. 5: (A) Maximum and Minimum values of Attributes, (B) Descriptive Values of Attributes.

In the exploration of the data of the profile corresponding to the coast (first profile), it is possible to determine the distribution of the values for the following attributes: latitude, precipitation, frequency, elevation (see Fig. 5 (B)). Standing out the fact that the elevation has a near behavior to a normal distribution in the complete range of the latitudes. On the other hand, the precipitation and the frequency have a exponentially falling distribution. The maximum and minimum values of these attributes can be seen in Fig. 5 (A).

7.3 Selection of Attributes

The criterion used to select attributes was to discard attributes with direct relationship to the geographic position. This pursues to avoid to introduce bias in the generation of models.

Given this, four attributes were selected: Frequency (F), Precipitation (P), Elevation (E), and Latitudinal Distance (LD).

8 Data Mining

8.1 Association Rules

For all the models, we considered 600 instances corresponding to the union of the decomposition of the three nearest to the coast profiles. Each profile was decomposed in 200 atoms. Table 2) summarizes association rules generated by using the WEKA [27] software tool. This rule indicates the presence of some associations among the attributes that could show some relationship between relief and climate. In this model interesting rules are generated that associate the Latitudinal Distance, the Frequencies and the Elevations in the first rules (we show only some rules that indicate the presence of relationships that we are looking for).

Table 2: Association rules found to relate attribute values.

Rule number	Antecedent	Consequent	Confidence
17.	F=middle-high LD=low E=high **17**	P=middle **15**	acc:(0.84262)
18.	F=middle E=middle-under **23**	LD=low **20**	acc:(0.83892)
19.	F=middle-high E=high **23**	P=middle **20**	acc:(0.83892)
20.	E=middle-under **123**	LD=low **101**	acc:(0.80862)
21.	F=middle **120**	LD=low **96**	acc:(0.78732)
22.	E=middle-high F=low **22**	P=low **18**	acc:(0.78646)
23.	E=middle-high LD=middle **16**	P=low **13**	acc:(0.7717)
24.	F=middle-under E=low **16**	LD=low P=high **13**	acc:(0.7717)
25.	E=middle-high **120**	LD=low **92**	acc:(0.75974)
26.	F=low E=high **23**	P=middle **18**	acc:(0.75466)
27.	P=middle **203**	LD=low **151**	acc:(0.74333)

8.2 Correlations among Profiles

To quantify possible lineal relationships among the attributes mentioned above, we computed the Pearson correlation coefficient for each one of the three first profiles separately, obtaining the following results:

- In Tables 3A, 3B, and 3C, we can observe that the lineal correlation between frequency and precipitation is growing among the profiles, i.e., the first profile has smaller correlation than the second and the third ones, in the same way that the second profile has smaller correlation than the third profile. It was expected that the results were inverse, i.e., that the correlation among precipitations and frequency were bigger in the first profile than the others, and it was falling as the profiles go into the mountain range, from the coast

to the intermediate depression. This because the available precipitation data are only associated to the coast (first profile).

- In Tables 3A, 3B, and 3C, we can also observe the existence of a medium negative correlation among elevation and precipitation. This correlation increases as the profiles go into the mountain range of the coast.

Table 3. Correlations of Profiles

First Profile (A)

LD	1			
F	-0,197	1		
P	-0,191	0,039	1	
E	0,208	-0,07	-0,396	1
	LD	F	P	E

Second Profile (B)

LD	1			
F	-0,203	1		
P	-0,197	0,07	1	
E	0,12	0,063	-0,386	1
	LD	F	P	E

Third Profile (C)

LD	1			
F	-0,2	1		
P	-0,275	0,337	1	
E	0,23	-0,28	-0,487	1
	LD	F	P	E

8.3 Maps to confirm the presence of tectonic and lithologic elements in low and medium range frequencies

It had been determined previously that the components of low frequency were associated to structural factors (basic tectonic configuration), as illustrated with an example in Fig. 6A and Fig. 6B. Then it was necessary to corroborate that the medium range frequencies are associated to types of rocks and geologic bodies. As shown in Fig. 6C, medium range frequencies can associate to bodies and types of rocks according to their size. To obtain these results we applied the technique of subtraction of maps of convolution of medium range elevations.

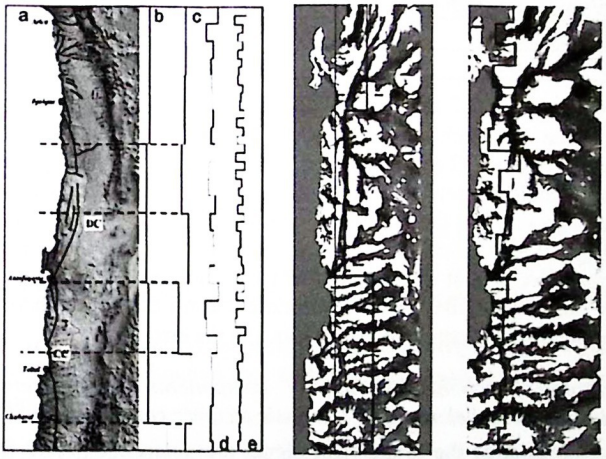


Fig. 6: A) General scheme [a) Relief, b) decomposition a1 (it bases) three blocks are appreciated along the whole region, c) decomposition d12 shows sections of the Fault of Atacama, the decompositions d10 (c) and d9 (d) you show elements lithologic of different magnitudes], B) Convolution Map, matching frequencies with faults, C)Convolution Map, matching frequencies with lithologic elements.

9 Interpretation and Evaluation

Based on the results obtained, we can make the following observations:

1. ***The existence of a relationship between relief and climate of latitudinal form in northern of Chile can not be determined in a precise way.*** **(R.1)**. This because the proposed models have results that conduct to contradictory conclusions.

   (a) On one hand, a contradiction takes place among the correlations of the profiles considering frequencies and precipitations (see Tables 3A, 3B, and 3C). It is obtained that the third profile (the farest one from the coast) has a correlation coefficient of *0,337*, which is bigger than the correlation obtained for the first profile (0,039) (the nearest one to the coast). Considering that the precipitation available data correspond only to the coast of northern of Chile, i.e., to the first profile, the corresponding correlation coefficient does not allow us to find evidence of the existence of lineal correlation among frequency and precipitation.

   (b) The association rules obtained indicate the existence of a relationship among a range of frequencies corresponding to the topography and the precipitations, but these rules have low support and confidence values (See Table 4).

Table 4. Association rules that relate Frequency and Precipitation values.

Antecedent	Consequent	Confidence
F=high 40	P=medium 26	acc:(0,635)
F=medium-high 44	P=medium 20	acc:(0,445)
F=medium 40	P=medium 17	acc:(0,416)
F=low 36	P=low 15	acc:(0,406)

   (c) Given the results R.2 and R.3 that are detailed to continuation, we eliminated of the space of search components of frequency associated to structural or lithological factors. The component range of frequency (high frequencies) remaining could contain information associated to the latitudinal variation of the precipitation. But the amplitude of these components of frequency are in the margin of the vertical error associated to the source of the topography. For this reason, we cannot discriminate which values can correspond to errors or to effect of the precipitations. This because doesn't exist yet a theoretical form to discriminate with confidence elevations in the error range.

2. ***We determine a relationship among components of frequency of the topography signal and elements of the basic tectonic configuration.*** **(R.2)**

   Three components of the topography signal are determined (low frequencies) associated to elements of the basic tectonic configuration. This through the decomposition via Wavelet and later on corroborated by STFT. An example of this can be appreciate in Fig.s 6B and 6C. We determined that the Wavelets decompositions corresponding to the blocks a12, d12 and d11 are related to structural factors. We associated those blocks with the latitudes corresponding to

each section, which correspondence of these elements was corroborated by geology experts by means of visual inspection. The verified structural elements correspond to the Atacama Fault in their three main segments: Salar del Carmen, Paposo, and Vicuña [21]. Also, we detected the presence of a block corresponding to inverse faults in the sector of the first region of Chile (Tarapaca).

3. ***We determine a relationship among components of frequency of the topography signal and lithology, together with geologic bodies in the study area.*** **(R.3)**

   A group of components of the topography signal is determined (medium range frequencies) that are associated to lithographs and geologic bodies at different scales. These bodies ranges from the 40 kms. to the 3.3 kms. of longitude approximately. They were detected using the Wavelet decompositions, which corresponds to the d10, d9, d8, d7, and d6 components. This was corroborated by using convolution maps and visual inspection of the correspondence of the components of frequency and the maps (see Fig. 6B and 6C).

4. ***We determined a medium-low negative correlation among the latitudinal variation of precipitations and the elevation of the relief.*** **(R.4)**

   When carrying out the correlation among the variation of the precipitation and the variations of the latitudinal form of the elevations, we obtain a correlation coefficient of -0,396 (see Tables 3A, 3B, and 3C). In spite of being a small value, it indicates a tendency that the geology expert considers interesting and that can be explored in later studies.

10 Discussion

10.1 Revision of the process

The process carried out in this study consisted mainly of two stages:

1. Generation of the data, which in turn is decomposed in the following steps:
   (a) First, we used GIS tools was used in order to generate from a DEM a longitudinal profile that corresponds to the topography signal in the study.

   (b) Later on, this signal is subjected to decomposition using Matching Pursuit (MP) with the Guimauve tool and decomposition via Wavelets with Matlab 7.0 tool.

   (c) With the data generated by the decomposition, we built a data table, in which were determined attributes that characterize the signal, and other additional attributes that characterize the relief and the precipitations.

2. Application of Data Mining techniques:

(a) With data table generated, it was carried out diverse discretizations on the attributes to be able to determine which one distributes the data to be studied in a better form.

(b) We carried out correlation tests among the different attributes to determine if they have some relationship type among them.

(c) We generated association rules using the algorithm PredictiveApriori algorithm implemented in the WEKA software tool, and whose results were analyzed and commented together with the results of the correlations.

This process and work can be improved in several forms. For example, we can improve the quantity and quality of the obtained data, as well as the use of other tools. In order to be able to discriminate the range of frequencies that currently are not totally analyze yet, we identify some changes or improvements that can be implemented in a following phase of this study:

1. The Guimauve tool carries out the decomposition of the signal through MP using a dictionary of STFT. However, this transformation may not be the most appropriate tool to carry out this work, given the asymmetric and acyclic characteristics of the signal. On the other hand, the Matlab software carries out the decomposition of the signal by using Wavelet, but their results are images, and it is not possible to determine quantitatively the frequency of the signal, which is very important for this study. Afterwards, the results obtained with Guimauve were graphically compared to the results obtained with Matlab. In the future, signal decomposition can be made by using software libraries, as alternative to the use of the software tools mentioned above.
2. For the data of the precipitations the available information corresponds to the average of the last 30 years of 8 coastal cities to carry out the interpolation. This restriction may introduce noise in the derived precipitation data, which in turn may affect the precision of the results. Therefore, it is necessary, in order to establish clearly the effects of precipitation over relief and vice versa, to have a bigger quantity of precipitation data, in particular coastal precipitation data and precipitation data associated to the coastal mountain range. This will allow us to carry out comparative studies between coastal mountain range and internal mountain range.
3. It is necessary to have a more precise map of rock types, since we used a simplified map that classifies the rocks for geologic time and not for geologic type. This will become necessary to be able to determine weather or not exist components of frequency associated to the rock type.
4. A parameter that may be tuned corresponds to the quantity of atoms in the decomposition of the signal using Guimauve. This because it may be feasible that a bigger quantity of atoms improve the results. Together with this, data discretization may affect the quality of the results obtained, and therefore it is an issue to be evaluated.

5. The lithological and structural components of the relief should be addressed in a future study, this with the purpose of minimizing the quantity of variables that may influence in the problem.

10.2 Expert Evaluation

The results obtained were presented to the expert, who made the following conclusions:

1. Apparently, the precipitations don't have relationship with the relief in the region under study (northern of Chile).
2. Given this, we can affirm that the climate has not varied significantly in the last 20 million of years. This conclusion is reinforced with studies of superficial datation in the area of the study. These studies date surfaces in northern Chile to that antiquity. This is related because if exists a bigger quantity of precipitations in the past these surfaces could not have been conserved. The sentence mentioned above is explained since if has existed in some moment inside the 20 million years a bigger quantity of precipitation, the landscape should not have been conserved. This is based on that to bigger quantity of precipitations, a bigger evolution of the landscape takes place.
3. This implies that the variations of the relief in the north area of Chile, you it is dominated by the basic tectonic configurations of the place and I didn't seize for the precipitations.
4. Is considered successful the carried out work for that although, it was not gotten to demonstrate the relationship between the precipitations and the relief. Data were gotten excellent about the factors that determine the relief. As well as it was proven that one can to work with ghastly analysis for analysis of topographies.

In summary, this work presents an application of the data mining process to the geologic domain, where data require transformation before applying them to the appropiate techniques. In this case, techniques require to transform the data to the frequency domain, where decomposition can be applied in order to identify components of the original signal. As shown, traditional statistical techniques (e.g., lineal correlation) and machine learning approaches (e.g., association rules) can be used to correlate and associate attributes values. In the case of association rules, attribute discretization must be applied before running the algorithm.

The user evaluation and interpretation of the results indicate that the relief in the zone under study (northern of Chile) is dominated mainly by tectonic variations and not for climatic variations.

11 Conclusions

The quantitative results obtained of the application of Matching Pursuit are not correlated for any constituent frequency with the latitudinal variations of the precipitations. However, other methods of establishing dependencies among data, specifically associations rules, show correlations among precipitation data and topographic latitudinal profiles. This is also corroborated with the visual results obtained of the application of the Wavelet Transform. Indeed, the decomposition with WT indicates us that the lithologic and tectonic factors are of first order in the control of the topography of the Coast Mountain in northern Chile. Thus, the constituent lower frequencies are controlled by the main segments of the AFS (Fig. 6 .A.a and 6 .A.b), while, those higher constituent frequencies are related with lithologic elements of diverse sizes (Fig. 6 .A.c and 6 .A.d). The highest frequencies can be associated to lithologic bodies until their amplitude (local relief) arrives to the margin of the vertical error. Therefore, the hypothetical existence of climatic control in the configuration of the relief of the Coast Mountain range, should be verified for morphologic features smaller than those that are possible to observe with GTOPO 90.

The application of this methodology allows us the study of factors involved in the evolution of the relief in a more analytic way together with outlining an approach based on the treatment of signals in the analysis of a relief. The combined application of the techniques of Transformed Wavelet and Matching Pursuit for the analysis of the relief allows us the identification of tectonic and lithologic features of the Coast Mountain of the northern Chile. Based on our results, we propose that the coastal relief between Arica and Caldera is fundamentally controlled by structural and lithologic features. Starting from the data of available precipitations, and the latitudinal gradients derived from those, it has not been possible to establish a latitudinal climatic control of the relief.

12 Limitations and Future Work

In terms of the methodology, a general revision of the should be carried out, with the purpose of determining whether the proposed steps are adequate, enough, or new tasks should be added.

In terms of the data, more precipitation data is required, in order to improve the interpolation task and the confidence of the correlation and association models.

In terms of the region under study, we wish to extend the study to other regions. For example, southern Chile where the quantity of precipitations (and the relief) are significantly different than the existent one in the northern Chile. This would allow us to validate and refine the proposed guide, and to contrast the results obtained and described in this paper.

In terms of the evaluation process, we want to incorporate a more formal method of evaluation of results, that complementing the expert domain evaluation, in way of making a more objective interpretation of the results.

In terms of the profiles generated and used in this study, we want to extend the study in order to determine the relevance of the existent relationship among the

lithography of an area and the formation of the relief in this area, by considering the latitudinal variations of the precipitations.

References

1. Bacry, E.: LastWave: The signal processing command language. Centre de Mathématiques Appliquees Ecole Polytechnique 91128 Palaiseau Cedex - France, (2004). http://www.cmap.polytechnique.fr/~bacry/LastWave/
2. Brachere, F.: Guimauve: GTK+ Matching Pursuit Decomposition. Observatoire Midi-Pyrenees Laboratoire d'astrophysique de Toulouse, France, (2002). http://webast.ast.obs-mip.fr/people/fbracher/
3. Burbank, D., Anderson, R.: Tectonic Geomorphology, Blackwell Science (2001).
4. Chapra, S., Canale, R.: Métodos numéricos para ingenieros, páginas 503-519, Tercera Edición. McGRAW-HILL,(1999).
5. Craig, C.: Vertical Accuracy Assessment and Archaeological Applications of 90m SRTM DEM in the Norte Chico Region of Coastal Peru, (2005).
6. Crave, A.: Quantification de l'Organisation des réseaux hydrographiques. Thèse Doctorale. Université de Rennes 1. France, (1995).
7. Cuartero, A., Felicísimo, A., Ariza, F.: Accuracy, Reliability, and Depuration of SPOT HRV and Terra ASTER Digital Elevation Models. IEEE Transactions On Geoscience and Remote, Vol 43, No2, February 2005.
8. Dirección General de Aeronáutica Civil, Dirección Meteorológica de Chile, (2005). http://www.meteochile.cl
9. Faundez, P., Fuentes, A., Reumay, P.: Procesamiento Digital de Señales Acusticas utilizando Wavelets. Instituto de Matematicas UACH, (2001).
10. Fayyad, U., Piatetsky-Shapiro, G., Smyth, P.: From Data Mining to Knowledge Discovery in Databases, (1996).
11. Glosario Geológico Vademécum REMER, (2005). http://www.proteccioncivil.org/vademecum/vdm02514.htm
12. Gribonval, R., Bacry, E., Abadia, J.: The stft package, Managing short time Fourier transforms and The mp package, Managing Matching Pursuit decompositions. IRISA, Campus de Beaulieu, Rennes Cedex, France, (2004).
13. Hartley, A., Chong, G.: A late Pliocene age for the Atacama Desert: Implications for the desertification of western South America. Geology, 30, 43-46, (2002).
14. Hernández, J., Ramírez, M., Ferri, C.: Introducción a la Minería de Datos. Capítulos 2 y 9. Pearson - Prentice Hall, (2004).
15. Jet Propulsion Laboratory: ASTER Higher-Level Product User Guide. Version 2.0. California Institute of Technology, (2001).
16. Labat, D.: Recent advances in wavelet analyses: Part 1. A review of concepts. Laboratoire des Mécanismes et Transferts en Géologie (LMTG), UMR 5563, CNRS/IRD/UPS, 14 Ave. E. Belin, Toulouse 31400. France, (2005).
17. Lague, D.: Dynamique de l'érosion continentale aux grandes échelles de temps et d'espace: modélisation expérimentale, numérique et théorique. Thèse Doctorale. Université de Rennes 1. France, (2001).
18. Mallat, S., Zhang, Z.: Matching Pursuit With Time-Frequency Dictionaries. Courant Institute of Mathematical Sciencies, Computer Science Departament, New York University, (1993).
19. Modelos digitales de elevación escala 1:50000. Generalidades y especificaciones, (2000). http://mapserver.inegi.gob.mx/geografia/espanol/normatividad /mde/menu.cfm? c=198

20. Montgomery, D., Balco, G., Willet, S.: Climate, tectonics, and the morphology of the Andes. Geological Society of America. v.29, n. 7, p. 579-582, (2001).
21. Quezada, J.: Interacción entre Procesos Endógenos y Exógenos en la Configuración del Relieve del Borde Costero del Anteraco del Norte de Chile. Technical Report. Universidad Católica del Norte. Chile, (2005).
22. Riquelme, R., Martinod, J., Hérail, G., Darrozes, J., Charrier, R.: A geomorphological approach to determining the Neogene to Recent tectonic deformation in the Coastal Cordillera of northern Chile (Atacama), (2002).
23. Unser, M., Aldroubi, A., Eden, M.: A Family of Polynomial Spline Wavelet Transforms. Biomedical Engineering and Instrumentation Program, National Centre for Research Resources. National Institutes of Health, Bethesda. USA, (1993).
24. Unser, M.: Splines: A Perfect Fit for Signal and Image Processing. IEEE Signal Processing Magazine, (1999).
25. Welch, R., Jordan, T., Lang, H., Murakami, H.: ASTER as a Source for Topographic Data en the Late 1990's. IEEE Transactions On Geoscience and Remote, Vol 36, No4, July 1998.
26. Wikipedia, (2005). http://es.wikipedia.org
27. Witten, I., Frank E.: Data Mining: Practical machine learning tools and techniques, 2nd Edition, Morgan Kaufmann, San Francisco, (2005).

Deriving Semantic Description Using Conceptual Schemas Embedded into a Geographic Context

Miguel Torres and Serguei Levachkine

Geoprocessing Laboratory – Centre for Computing Research – IPN, Mexico City, MEXICO
{mtorres, sergei}@cic.ipn.mx

Abstract. Information integration and semantic heterogeneity are not trivial tasks. An integrated view must be able to describe various heterogeneous data sources and their interrelation to obtain shared conceptualizations. In this paper, we propose an approach to make a conceptualization of the real world based on conceptual schemas, which are used to generate a semantic description of the geospatial domain. This description attempts to provide the guidelines to formalize the geographic domain in form of geospatial ontologies according to specific contexts. In this case, we propose that conceptual schemas are built in order to abstract specific and essential *parts* of the geospatial domain and to represent schematically what geospatial entities should be collected and how they must be organized. We perceive that geographic data modeling requires models that are more specific and capable of capturing the *semantics* of geospatial data, offering higher abstraction mechanisms and implementation independence. Therefore, we approach conceptual schemas to describe the contents of the real world abstraction to specify the behavior of the geospatial entities, in which the context plays an important role to guarantee shared and explicit conceptualizations. Our research is mainly oriented to propose an approach related to conceptual issues concerning what would be required to establish ontologies of the geospatial domain. In addition, the work is led to formalize appropriate methods to represent ontologies of the geospatial domain.

1 Introduction

Ontology has gained increased attention among researchers in geographic information science in recent years. Up-to-date, the ontology notion plays an important role in establishing robust theoretical foundations for geographic information science [1]. Under this umbrella, it is possible to unify several interrelated research subfields, each of which deals with different perspectives on geospatial ontologies and their roles in geographic information science. Three broad sets of foundational issues need to be resolved: (1) conceptual issues concerning what would be required to establish an exhaustive ontology of the geospatial domain, (2) representational and logical issues relating to the choice of appropriate methods for formalizing ontologies, and (3) issues of implementation regarding the ways in which ontology ought to influence the design of information systems.

Nowadays, there are diverse institutions that use geospatial data to make a decision in different environments. The use of geographic databases through geographic information systems (GIS) provides tools for managing, analyzing and processing geo-

© M. Moreno-Ibarra, S. Levachkine (Eds.)
Semantics in Geoinformatics Applications
Research in Computing Science 25, 2006, pp. 109-119

spatial data. However, information can not sometimes be represented in "adequate" way, since it presents ambiguities that do not allow the appropriate use and analysis. These ambiguities are originated by imprecision of information, heterogeneity and isolation sources. Whereby, it is difficult to develop interoperable applications that allow us to share, integrate and represent geospatial information.

These facts bear with searching solutions oriented to geospatial data representation and integration, semantic heterogeneity and imprecise geographic objects issues. Consequently, commercial GISs do not have tools to extensible explore the essential properties and relations of geographic objects. Therefore, by means of these applications, it is difficult to explore the semantics of a set of geographic objects.

According to [2] and [3], the ontologies and the knowledge representation are essential for the creation and use of standards to exchange data, as well as for the design of human computer interaction, whereby an ontology allows us to solve problems associated to heterogeneity, interoperability, representation, integration and exchange of geospatial data. These problems imply incompatibility between diverse geographic objects, as well as a different spatial conceptualization according to a specific context. For example, we engage with the world day by day in a variety of different ways: we use maps, specialized languages, and scientific instruments; we also engage in rituals and tell stories; we use information systems, databases different machines and other software-driven devices of various types. Each of these ways of engaging with the world, we shall now say, involves a certain *conceptualization.* What this means is that it involves a system of concepts and categories, which divide up the corresponding universe of discourse into objects, processes and relations in different sorts of ways. Thus, in a religious ritual setting, we might use concepts such as *God*, *salvation* and *sin*; in a scientific setting, we might use concepts such as *micron*, *force* and *nitrous oxide*; in a story-telling we might use concepts such as *magic spell*, *dungeon* and *witch*. These conceptualizations are often tacit, that is, they are often invisible components of our cognitive apparatus, which are not specified or thematized in any systematic way [4].

On the other hand, the conceptualization of geospatial domain is diverse, because the geospatial data used are often imprecise or many subjects have different point of view. Thereby it is important to consider alternative object representations, which are independent of the imprecise nature of the geospatial data [5].

Our research is mainly oriented to propose an approach related to conceptual issues concerning what would be required to establish ontologies of the geospatial domain.

In this paper, we propose an approach to make a conceptualization of the real world based on conceptual schemas, which are used to generate a semantic description of the geospatial domain. This description can provide the guidelines to formalize this domain in form of geospatial ontologies according to specific contexts.

The rest of the paper is organized as follows. Section 2 describes some related works. In Section 3, we describe the proposed approach to conceptualize the geospatial domain. Section 4 shows a case study to build a semantic description based on conceptual schemas. Our conclusions and future works are outlined in Section 5.

2 Related works

Some works related to ontologies and semantics in geospatial information science to be mentioned are as follows. Guarino [6] coined the term "ontology-driven information systems" and provided a broad discussion on their place in the computer and information science. Gruber, one of the pioneers of the use of ontological methods in information science, defines an ontology as "a specification of a conceptualization" in [7].

Smith *et al.* [4] reported the results of a series of experiments designed to establish how non-expert subjects conceptualize geospatial phenomena. Subjects were asked to give examples of geographical categories in response to a series of differently phrased elicitations. The results yield an ontology of geographical categories – a catalogue of the prime geospatial concepts and categories shared in common by human subjects independently of their exposure to scientific geography.

Bishr *et al.* [8] argued that information modeling requires to be controlled to allow successful sharing of information. Also, they suggest that any coherent information model need to be based on accepted ontological foundation to guarantee unambiguous interpretation. In addition, their work attempts to show that ontology based information modeling provides more cognitive foundation for information systems models and therefore minimizes the problem of semantic heterogeneity.

Smith *et al.* [2] designed an ontology of geographic kinds to yield a better understanding of the structure of the geographic world, and to support the development of GIS that are conceptually sound. This work first demonstrated that geographical objects and kinds are not just larger versions of the everyday objects and kinds previously studied in cognitive science.

Fonseca *et al.* [9] proposed a framework to link the formal representation of semantics (i.e., ontologies) to conceptual schemas describing information stored in databases. The main result is a formal framework that explains the mapping between a spatial ontology and a geographic conceptual schema. The mapping of ontologies to conceptual schemas is made using three different levels of abstraction: formal, domain and application levels. At the formal level, highly abstract concepts are used to express the schema and the ontologies. At the domain level, the schema is regarded as an instance of a generic data model. At the application level, authors focus on the particular case of geographic applications. Additionally, they discuss the influence of ontologies in both the traditional and the geographic systems development methodologies, with an emphasis on the conceptual design phase.

According to this works and in particular with [9], it is important to distinguish that our research is concentrated to use conceptual schemas to describe the semantic contents of the real world abstraction to specify the behavior of geospatial entities, in which the context plays an important role to guarantee shared and explicit conceptualizations. We will aim to propose issues and methods concerning what would be required to establish and to represent ontologies of the geospatial domain.

3 Geospatial domain conceptualization

This section gives the guidelines to build conceptual schemas to conceptualize the geospatial domain. Thus conceptual schemas are used to generate a semantic description, which can provide the framework to formalize the geospatial domain, according to specific contexts. In this section, we point out the most important components involved in our approach such as conceptual schema and context.

3.1 How to design conceptual schemas for geospatial domain

In the traditional systems modeling approach, the modeler is required to capture a user's view of the real world in a formal conceptual model. Such an approach forces the modeler to mentally map concepts acquired from the real world to instances of abstractions available in his paradigm choice. On the other hand, the consolidation of concepts and knowledge represented by a conceptual schema can be useful in the initial steps of ontology construction. To adequately represent the geographic world, we must have computer representations capable not only of capturing descriptive attributes about its concepts, but also of describing the relations and properties of these concepts.

We propose conceptual schemas to describe the contents of the real world abstraction in order to specify the behavior of the geospatial entities. In this case, conceptual schemas certainly correspond to a level of knowledge formalization. In this case, conceptual schemas are built to abstract specific parts of the geospatial domain and to represent schematically what geographic entities should be collected and how it must be organized. We perceive that geographic data modeling requires models that are more specific and capable of capturing the *semantics* of geospatial data offering higher abstraction mechanisms and implementation independence.

The proposed conceptual schemas are composed of two types of concepts (C): *terminal* (C_T) and *non-terminal* (C_N). The first ones are concepts that do not use other concepts to define their meaning (they are defined by "simple values"). The meaning of non-terminal concepts is conceived by other concepts, which can be terminal or non-terminal concepts (see Eqn. 1).

$$C = C_N \bigcup C_T \tag{1}$$

Each concept has a set of *aspects*. Aspects are characteristics that describe the properties, relations and instances that involve the geospatial objects. From-now-on, we shall use the term "relation" to denote unary relations/properties as Berendt *et al.* [10]. From this point of view, all aspects of a terminal concept are simple, e.g. the type of all aspects that belongs to the set of primitive types (punctual, linear and areal objects) is denoted by (T_P), as shown in Eqn. 2.

$$\begin{aligned} T_P &= \{number, character, string, enumeration, struct\}, \\ A &= \{a_i \mid type(a_i) \in T_P\}, \end{aligned} \tag{2}$$

where T_P is the set of primitive types; A is the set of aspects.

Then, the set of *terminal concepts* is defined by Eqn. 3.

$$C_T = \{c(a_1, a_2, ..., a_n) \ni a_i \in A,\ i = 1,..,n\} \tag{3}$$

In the same way, the *non-terminal concepts* have at least one aspect that does not belong to T_P. It is denoted by Eqn. 4.

$$C_N = \{c(a_1, a_2, ..., a_n) \ni \exists a_i \notin A\}, \text{ where } c \text{ is a concept.} \tag{4}$$

Finally, the set of relations R is defined by the pairs that are associated to Γ and Φ, in which Γ and Φ are non-reflexive, non-symmetric, and transitive relations (Eqn. 5).

$$R = R_\Gamma \cup R_\Phi = \{(a,b) \mid a \Gamma b,\ a \in C_N,\ b \in C\} \cup \{(a,b) \mid a \Phi b,\ a \in C_N,\ b \in C\} \tag{5}$$

According to definitions presented above, it is necessary to express the semantics that can provide a conceptual schema by means of a description D. Therefore, we consider the concepts C embedded into the conceptual schemas through geospatial objects, which are represented by primitive types as well as the set of relations R involved among geospatial objects (see Eqn. 6)

$$D = \langle C, R \rangle \tag{6}$$

Fig. 1 depicts a conceptual schema, which has been designed for the geospatial domain. Also, this schema is adaptive for any context. In other words, it attempts to reflect the main features involved in this domain. For instance, if we have topographic, tourism, or geologic contexts, it is possible to describe the entities, characteristics and relations embedded between geographic objects. The main features involved into geospatial domain have been abstracted of the real world in order to obtain a conceptualization. This conceptualization provides us an explicit vocabulary that represents the ontological commitment of the cognitive and intuitive perception of the subjects.

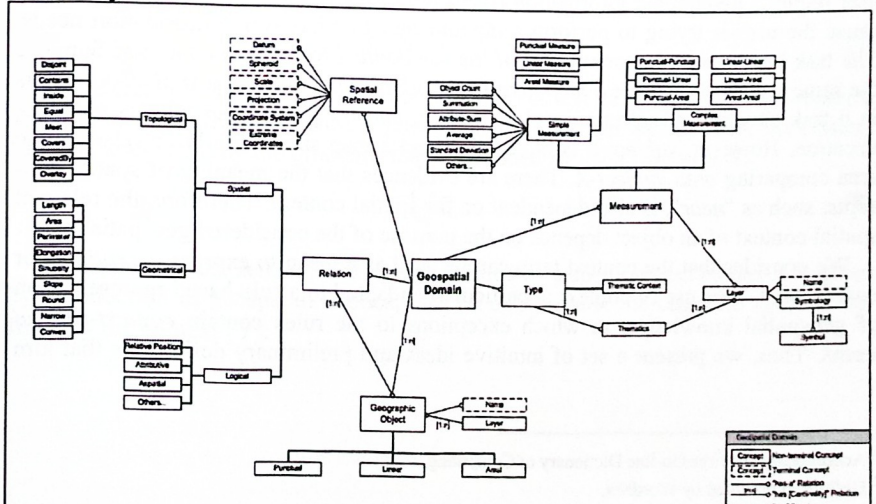


Fig. 1. Conceptual schema proposed to conceptualize the geospatial domain

The conceptual schema shown in Fig. 1 conceptualizes the geospatial domain. This schema represents a raw conceptualization, which contains an optimal number of relations. We are looking for a compact conceptual schema-based abstraction that drives the cognitive process of phenomenon semantic description under specific context. We consider that conceptual schema depicted in Fig. 1, could have more concepts involved in geospatial domain, thereby it can be a certain approximation about the main entities that compose this domain. This domain could be considered as a first step to collect and organize the concepts contained into the geographic context.

3.2 Context-driven approach to restrict the domain into conceptual schemas

The context term is defined as *"that which surrounds, and gives meaning to something else*[1]*"* or it is the *"discourse that surrounds a language unit and helps to determine its interpretation*[2]*"*. However, to obtain shared conceptualizations and to accomplish with ontological commitments, it is necessary to take into account the *context* term. Also, this term can be used to consistently map different conceptualizations. Due to this, the meaning of a spatial concept may be dependent on a large number of contexts within which the concept is used.

Contexts about a particular use of a spatial concept refer to the knowledge that human uses to constrain the meaning of communication. To reach a common understanding of a vague concept, e.g. ***near***, the system and the user require to share knowledge about the relevant contexts that affect the understanding of the vague concept. Among many potential contexts that may affect how people understand spatial concepts. We focus on three of them: *task*, *spatial contexts*, and *background of the user*. We perceive that context is a key issue in interaction between human and computer, describing the surrounding facts that add meaning.

Context is very useful in geographic information science. For instance, when a user requests geographic information (map) to be displayed by a GIS, it is often because the user is trying to perform a *domain task* that has some information needs. The task becomes an important part of the use *context* for *spatial concepts*. Suppose the same request *"show me a map near Cancun"* may be made by a *subject-A,* who is in a task situation of selecting a clothes shop, and by a *subject-B* who is planning vacation. However, *subject-B* is likely to expect a map showing a larger geographic area comparing with *subject-A*. There are evidences that the meaning of spatial concepts, such as *"near"*, is also dependent on the spatial context. Therefore, the relevant spatial context of an object depends on the purpose of the considered geospatial data.

We consider that the context term can be used as a mean to express *exceptions*[3] or *constraints*[4]. This use of context is particularly adapted to a rule-based representation of geospatial knowledge, in which exceptions to the rules contain *context-related* terms. Thus, we present a set of intuitive ideas and preliminary definitions that aim

---

[1] According to the Free On-line Dictionary of Computing.

[2] Definition provided by WordNet.

[3] Example: "remove all buildings except the isolated one".

[4] Example: "the river must be into a valley".

better understanding the roles that play the context into the conceptualization based on conceptual schema.

- **Context.** Let a set (of terminal and non-terminal concepts) X, which contains a set of subcontexts Y and $X \subseteq Y$. Then, the set of subcontexts composes the universe of the context denoted by $Y \subseteq C_G$, in which C_G is called *geographic context.* X should be a large set ("large" with respect to cardinality $|C_G|$). Thus,
  - A concept C, which can be terminal (C_T) or non-terminal (C_N) concept, belonging to subcontext Y should mentally suggest or bring into our attention Y.
  - $C \in Y$ implies that the name (mention, evocation[5]) of C makes us to think of Y. In the real world, Y{occurs, appears, is produced, is achieved, happens, is used} whenever C {occurs, appears, is produced, is achieved, happens, is used}. For example, concept river ∈ context HYDROLOGY.HYDROLOGY is a set, but we wrote here just its name, since it is a named set.
  - Context should be obvious, not hidden. It should be evoked by every C belonging to it.
  - Context is the extension of *concept* to sets (to named sets).
  - A concept may belong to several contexts. For example, river ∈ HYDROLOGY, river ∈ WATER FLOW. A concept (belonging to a context) could be a context, too. For example, MEXICAN HYDROLOGY ∈ HYDROLOGY.
  - Contexts can overlap.

In conclusion, we propose additional intuitive definitions, which give us ideas to describe the context.

- *Problem* or *Objective* (P). It contains initial state and ending state, in other words, the study object (O_i), a result object (O_r) and a set of constraints (K) that involve the problem or objective (see Eqn. 7).

$$P_m = \{O_i, O_f, K_m\}, \quad K \rightarrow K_m \tag{7}$$

Therefore, we should take into account the context of the problem to obtain a shared conceptualization about the phenomenon of the real world. Then, the context (Ψ) can be denoted by the problems that are defined by itself (see Eqn. 8).

$$\psi = \bigcup_m P_m \tag{8}$$

In conclusion, the semantics is always defined by a specific context and it is given by a collection of geospatial entities, thus an entity inside the semantic space is defined by the context (Ψ).

[5] Thinking, depicting, imagining.

4 Case study

In this section, we describe two scenarios, which are focused on showing how to conceptualize the geospatial domain, by means of conceptual schemas in order to obtain a semantic description regarding specific context. The goal is to depict how these scenarios converge in the same semantic description (see Fig. 4). Although their representations are different, they belong to the same context; thereby their semantic description is the same as well as their conceptualization.

- ***Scenario 1: Imagining the real world.*** Suppose that we are seeing a landscape, which depicts several entities such as a forest that has a lake and a river. Moreover, the freeway F25 crosses the highway I37, F25 is used to arrive to Santa Cruz that is the main town of the surroundings (see[6] Fig. 2). So, it is important to make a conceptualization about our observations. In other words, we are making and abstraction process that is used to conceptualize the landscape, and then this kind of conceptualization can be represented in a conceptual schema and restricted by a context. We use the conceptual schema described in Fig. 1 to generate the semantic description.

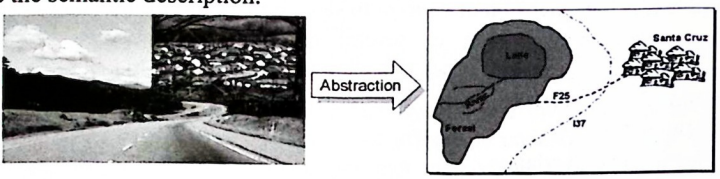


Fig. 2. Scenario 1: Imagining and representing the real world

- ***Scenario 2: Vector map.*** Suppose that we are seeing a map (Fig. 3), it depicts different thematics that consist of different layers, in which each layer contains geographic objects represented by geospatial primitives. The map has *Populations* (POP), *Hydrologic Features* (HYF), *Roads* (ROD) and *Soils* (SOL). Additionally, each thematic and its layers are denoted by a legend. The map is composed of 2 areal objects, 3 linear objects and 1 punctual object.

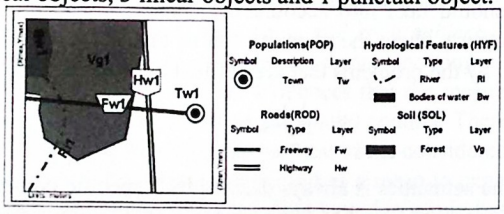


Fig. 3. Thematic map used to illustrate the second scenario

Thereinafter, we use the conceptual schema shown in Fig. 1 to describe both scenarios. According to Fig. 1, in order to obtain a semantic description from conceptual schema, it is necessary to map the geospatial entities into the conceptual schema. Once concepts have been defined into the conceptual schema, we choose the non-

[6] Fig. 3 is only an approximation or an idea of a general landscape described above, we only select some objects to show an illustrative example.

terminal concept to be described (this means to select the aspect to be pointed out). The process continues until we find a terminal concept. When the terminal concept is found, it is necessary to select a pair of geospatial objects, verifying whether a relationship between them exists, otherwise a part of description needs to be generated. Terminal concepts are defined by the kind of relation between two objects. In other words, the description starts at the non-terminal concept called ***"Geospatial Domain"***. Also, the non-terminal concepts are denoted by means of *rectangles* and the values of the terminal concepts are represented by *ellipses*.

According to the aspect of each non-terminal node, we establish a relation that defines another non-terminal or terminal concept (depending on the objective). This leads to complete the description of geographic objects that compose both scenarios. They converge in the same description according to the context; even though these are represented in different ways (see Fig. 4).

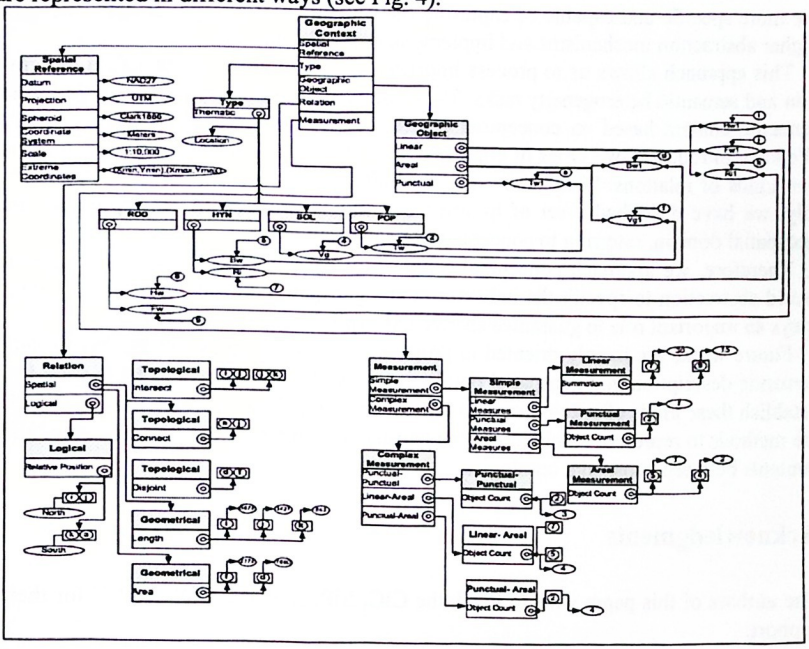


Fig. 4. Semantic description of the scenarios 1 and 2

The method is focused on describing the *semantic content* based on conceptual schemas in a geographic context. However, the description depends on a number of relations, properties and measurements[7] that are considered, whereby it is possible to increase the semantic resolution in the description. The description is made using *tuples* of non-terminal and terminal concepts related among themselves (they are denoted by *Concept* ***relation*** *Concept*). For instance, Fig. 2 and 3 are composed of

[7] A measurement is a procedure for computing values, which are the basis to evaluate characteristics of geospatial phenomena.

several spatial objects. The objects in the layer reflect the relation **"is-a"** (i.e., *HW* **is-a** *Linear Object*). Moreover, the topological relation ***"Intersect"*** is related to *Hw1* and *Fw1*, which both are linear objects. Thus, in description the ***"Intersect"*** relation is generalized as a spatial relation too.

5 Conclusion and future works

We propose an approach to make a conceptualization of the real world based on conceptual schemas, which are used to generate a semantic description of the geospatial domain. This description attempts to provide the guidelines to formalize the geographic domain in form of geospatial ontologies according to specific contexts.

On the other hand, we perceive that geographic data modeling requires models that are more specific and capable of capturing the semantics of geospatial data, offering higher abstraction mechanisms and implementation independence.

This approach allows us to process imprecise data and aid to information integration and semantic heterogeneity tasks. Thus, the method is focused on describing the semantic content based on conceptual schemas embedded into geographic context. We have introduced two types of concepts: "terminal" and "non-terminal" as well as two kinds of relations: *"has"* and *"is-a"* to build the conceptual schema. Additionally, we have described a set of intuitive definitions oriented to conceptualize the geospatial domain, referring to conceptual schemas and context.

Therefore, we approach conceptual schemas to describe the contents of the real world abstraction to specify the behavior of the geospatial entities, in which context plays an important role to guarantee shared and explicit conceptualizations.

Future works are mainly oriented to propose conceptual issues related to translate semantic description into geospatial ontologies, as well as what would be required to establish these kinds of ontologies. In addition, our work is led to formalize appropriate methods to represent ontologies of the geospatial domain and to measure semantic contents between geospatial ontologies.

Acknowledgments

The authors of this paper wish to thank the CIC, SIP, IPN and CONACYT for their support.

References

1. Mark, D., Smith, B., Egenhofer, M. and Hirtle, S.: Ontological Foundations for Geographic Information Science, in McMaster, R. and Usery, L. (Eds.) *A Research Agenda for Geographic Information Science*, CRC Press, Boca Raton, FL (2004) 335-350.
2. Smith, B. and Mark, D.: Ontology and Geographic Kinds. *Proceedings of the 8th International Symposium on Spatial Data Handling*, Vancouver, Canada (1998) 308-320.
3. Minsky, M.: *A Framework for Representing Knowledge*, Technical Report, in MIT-AI Laboratory, AIM-306, USA (1974).
4. Smith, B. and Mark, D.M.: Geographical categories: an ontological investigation. *International Journal of Geographic Information Science*. 15(7) (2001) 591-612.

5. Torres, M., Moreno, M., Quintero, R. and Fonseca F.: Ontology-driven description of spatial data for their semantic processing. *Proceedings of the First International Conference on Geospatial Semantics*, Springer-Verlag, 3799, Mexico City, Mexico (2005) 242-249.
6. Guarino, N.: Formal Ontology and Information Systems. *Proceedings of the International Conference on Formal Ontology in Information Systems*, Kluwer Academic Publishers, IOS Press, Trento, Italy (1998) 3-15.
7. Gruber, T.R.: Toward Principles for the Design of Ontologies Used for Knowledge Sharing. *International Journal of Human and Computer Studies*. 43(5/6) (1995) 907-928.
8. Bishr, Y. and Kuhn, W.: Ontology-Based Modelling of Geospatial Information. *Proceedings of 3rd AGILE Conference on Geographic Information Science*, Helsinki, Finland (2000) 24-27.
9. Fonseca, F., Davis, C. and Câmara, G.: Bridging Ontologies and Conceptual Schemas in Geographic Information Integration. *GeoInformatica*. 7(4) (2003) 355-378.
10. Berendt, B., Barkowsky, T., Freksa, C. and Kelter, S. I. E.: Spatial Cognition - An Interdisciplinary Approach to Representing and Processing Spatial Knowledge. *Spatial Representation with Aspect Maps*, Springer-Verlag, (1998) 157-175.

Spatial Analysis to Generate a Conceptualization of Geometric Properties

Karina Verastegui, Marco Moreno-Ibarra and Miguel Torres
Geoprocessing Laboratory-CIC- National Polytechnic Institute, Mexico City, Mexico
verastegui@sagitario.cic.ipn.mx,
{marcomoreno, mtorres}@cic.ipn.mx
http://geo.cic.ipn.mx

Abstract. We propose an approach oriented to generate an explicit representation of the geometric characteristics of geospatial data. Our methodology is based on a conceptualization of a geospatial domain. The concepts are extracted automatically by analyzing the properties of geospatial objects Also the concepts are stored in the spatial database to support subsequent processing. The conceptualization of geometric properties is based on measurements of geospatial objects. The results of this measurement are classified to obtain representative clusters of values in order to describe these properties. The values are used to define which concept better represents the properties of each object. We consider different geometric characteristics in a particular case study, such as rivers. Then, the characteristics that we have considered are the following: sinuosity, length, slope, and course. By using this method, we attempt to catch the semantics that contains geospatial data in the geometric context.

1 Introduction

This research has as main purpose to design a Geographic Information System (GIS) oriented to explicitly represent the geometry of geospatial data. The explicit representation is based on the conceptualization of the geometric properties. This is represented by means of concepts that describe different geometry. To quantify the characteristics, different measurements are computed to make a conceptualization.

In this paper, we present a conceptualization oriented to generate ontological information and thus structure or homogenize the databases. We are looking for a format of data representation, which does not depend on scale, format, references, etc. According to [2] a body of formally represented knowledge is based on a *conceptualization*: the objects, concepts, and other entities that exist in some area of interest and the relationships that hold among them. A conceptualization is an abstract, simplified view of the world that we wish to represent for some purpose [3].

Ontologies encode semantic relations between concepts and hence facilitate the detection of associations between related terms. In modern information systems based on ontologies, one seeks canonical descriptions of knowledge domains and associated classificatory theories.

Traditionally, geographic data models usually explicitly represent a set of basic objects, their geometry and their properties. But much of the geographic world's semantics appears in the relations linking objects [12] [8]. Nevertheless, most of these rela-

© M. Moreno-Ibarra, S. Levachkine (Eds.)
Semantics in Geoinformatics Applications
Research in Computing Science 25, 2006, pp. 121-133

tions are not explicitly represented in data models describing geographic databases. Usually, these relations only implicitly appear when one is looking at a *display* of a geographic database [7]. Geographic Information Systems (GIS) handle geo-spatial data at different levels of details. Due to page limit our literature overviews stopped at this point.

In this work, we generate a conceptualization of geometrical properties of geospatial objects and explicitly represent them into spatial databases. The paper is organized as follows: Section 2 describes the measurements to quantify the geometric characteristics. Additionally, Section 3 points out the conceptualization method. Some preliminary results are shown in Section 4, and Section 5 sketches out our conclusion and future work.

2 Conceptualization of geometrical characteristics

This research has as main purpose to design a Geographic Information System (GIS) oriented to explicitly represent the geometry the geospatial data. The explicit representation is based on the conceptualization of the geometrical properties (See Fig.1). The conceptualization of geometrical properties is based on measurements of geospatial objects. The values obtained by the measurements are classified to obtain representative clusters of values in order to describe these properties. The values are used to define which concept better represents the properties of each object.

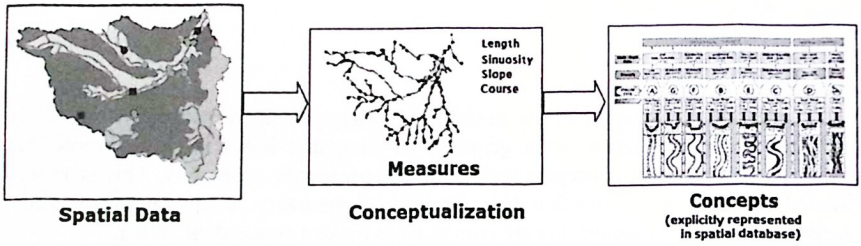


Fig. 1. Process to generate concepts explicitly represented

2.1 Measures

The necessity to measure is evident in most of the technical or scientific activities. Nevertheless, it does not only interest to count on measurements, but also to know if these measurements are valid. The validity of the measurement in any technical or scientific discipline is based on the respect to the principles of the general theory of the measurement (in particular, we will lean in the call representational theory of the measurement). The foundation of the representational theory consists of which all measurement must assure a suitable representation of the real attribute measured by

means of the symbols or assigned numbers. Additionally, the relationships should be preserved among this entity.

In order to establish measurements, we must divide our observation of the real world or dominion. We must identify which are the organizations that we want to measure (e.g., river) and to define what attribute we wished to characterize (e.g., the sinuosity). The measurement assigns a value to each organization to characterize its attribute (e.g., the river S3 has a sinuosity of 2.75). The most important is that the measurement that we establish is not inconsistent with the relations observed in the real world.

It is important to consider that not always the ideas on the attributes are so clear. We can begin by simple subjective valuations (e.g., to use questionnaires where they are classified or orders the opinions of the experts on an attribute), that do not constitute measurements from the point of view of the theory of the representation, but that they can be analyzed to improve the understanding on the real world. It is possible that after accumulating data of this type, it is necessary to be obtained to define a formal measurement. An allocation that settles down between real world and values of measurement usually is denominated measurement scale.

To study with detail the geometry of the spatial objects different simple algorithms should be developed, in order to reflect in some aspects the basic characteristics of the geometry. For example, in the generalization of a hydrological network some such geometric problems exist as lines that you angular and they lose detail, as well as nodes that are displaced as is shown in Fig. 2. In the generalization of a map, it takes to those contradictions exist between the demands of geometric accuracy and the geometric coincidence. The geometric exactitude of a map presupposes that each object of the terrestrial surface is exactly represented in the map in the same place with its true plane contours and dimensions, conserving the positions of all the objects and the distances that overcome them, correctly of agreement with the scale of the map.

Fig. 2. Generalization of hydrological network

This reason is important to analyze, the change that occurs in the geometric form in to the objects. A form to make this process is by means of the measurement. In other works, this is the base of value the geometric characteristics in the generalization. In the Technical Annex of the project AGENT a classification of geometric measurements is shown and it is focused on the generalization. With this classifica-

tion will be carried out measurements of the geographical objects with the purpose of obtaining a conceptualization, guided to evaluate the result of the generalization.

In conclusion the measurements are numeric values assigned to an observation that reflects a magnitude, quantity or a characteristic. For this reason, we can make measurements according to their geometric form.

3 Methodology

The explicit representation is based on the conceptualization of the geometrical properties. This method does not depend on the scale as the traditional GIS approaches. In the following subsections, we outline our methodology.

3.1 Methodology of geometrical conceptualization

In our model, the geometry of geospatial objects will be described by means of concepts that represent geometrical properties, such as size, sinuosity / complexity, elongation / eccentricity, compactness, as well as other important aspects. We consider three types of geospatial data; point-like objects (i.e. well or tree), linear objects (i.e. a road or river) and the area objects (i.e. state boundary or lake).

The conceptualization process consists of four steps: (1) to obtain the measurements, we use algorithms that evaluate geometric characteristics. Fig. 3 shows a classification of measurements. Shape describes the geometric representation of spatial objects according to [11].

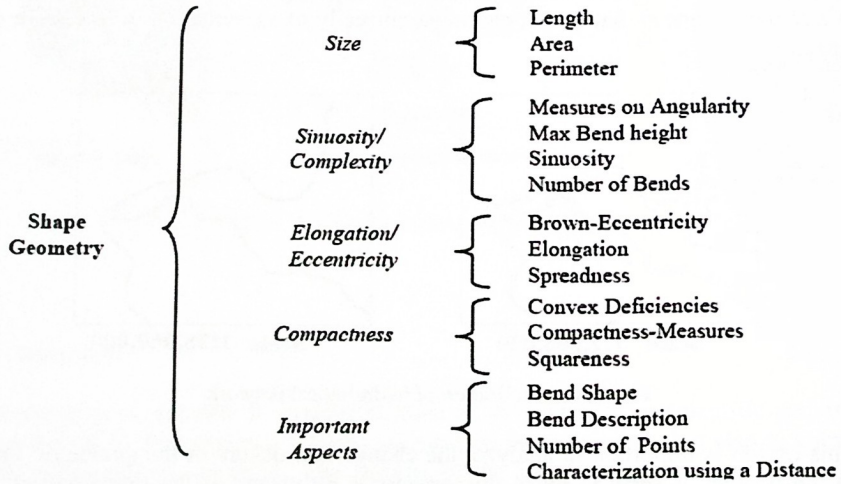
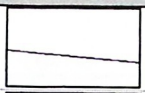
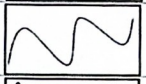


Fig. 3. Classification of Measures

(2) The obtained values by the measurements are classified to generate representative clusters of values in order to describe the geometric characteristics in a qualitative way (3) each cluster will be used to assign a concept according to the classification of intervals of the measurements (see Table 1). Fig. 4 depicts a workflow diagram of the conceptualization process.

Table 1. Mapping between the obtained values of a measurement and the concept

Spatial representation	Range (measure)	Concept
	1 - 1.4	Straight
	1.5 – 2.0	Little sinuous
	2.1 – 2.4	Half sinuous
	2.5 – 3.5	Very sinuous

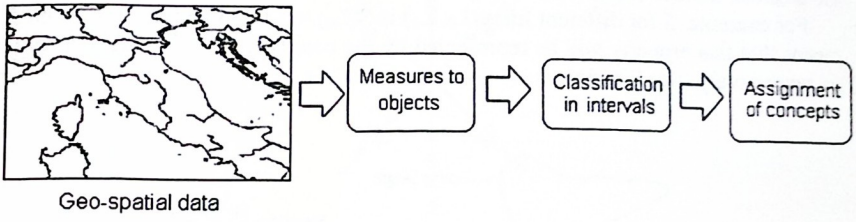


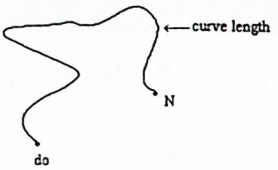
Fig. 4. Workflow diagram to obtain the conceptualization

3.2 Methodology of geometrical conceptualization of rivers

In this investigation was carried out the conceptualization of rivers by means of measurements (e.g. length, sinuosity, heights, trajectory and address of rivers), the values obtained by the measurements are classified in intervals, assignment of concepts according to the classification of intervals of the measurements.

3.2.1 Compute of length measurements

The length measurement (L), it is obtained by the equation 1, which is computed by the shortest distance between a point and a straight line is perpendicular one to this straight line that goes to that point. In other words, if we think of vectors instead of lines, it would be the vector that goes to the point of measurement of the distance that maintains a product to climb = 0.



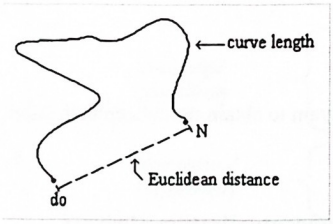
$$d_{(P,L)} = \frac{(y_0 - y_1)x + (x_1 - x_0)y + (x_0 y_1 - x_1 y_0)}{\sqrt{(x_1 - x_0)^2 + (y_1 - y_0)^2}}$$

Fig. 5. Compute of length measure

3.2.2 Compute of sinuosity measurement

In the case of *sinuosity measurement* (*S*), it is obtained by the equation 2, which is computed by dividing curve length with Euclidean distance from the initial node d_0 to the terminal node N ($P(d_0,N)$) (See Fig. 6a).

For example, *S* for different lines (L_0, L_1) is $S(L_0) = 1.0034$ and $S(L_1) = 3.3905$. It means that this property will be represented by the concepts *straight* and *very sinuosity* respectively (See Fig. 6b).



$$Sinuosity = \frac{Curve\ length}{P(d_0, N)}$$

a)

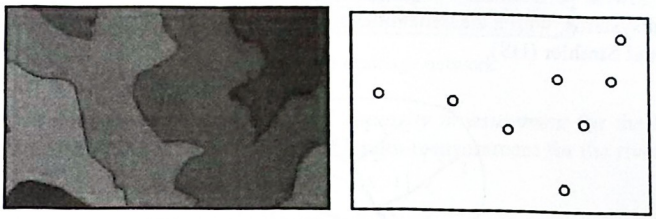
Classification of S	Concept
1.0 – 1.4	Straight
1.5 – 2.0	little sinuous
2.1 – 2.4	half sinuous
2.5 – 3.5	very sinuous

b)

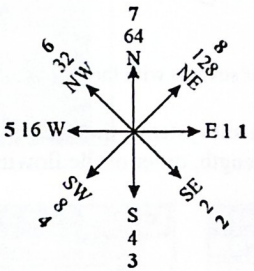
Fig. 6. Compute of sinuosity measurement

3.2.3 Compute of heights, trajectory and address of rivers Measure

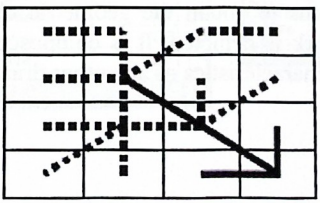
The slope of the land is computed by starting from the gradient of heights among neighboring cells (Fig 7a). The line of flows describes the address of the glides of water in the surface (Fig 7b), following the trajectory of more slope. To code the address of the glide, a code is assigned to each cell, among a total of 8 different values; each value represents the orientation of the cell, following the sketch: E=1, SE=2, S=4, SW=8, W=16, NW=32, N=64, NE=128 (Fig. 7c).



a) Compute of heights



c) Direction of flood



b) Compute of trajectory

Fig. 7. Compute of heights, trajectory and address of rivers

3.2.4 Classification in subsystems

To carry out a classification in everything the one laughed it is necessary to make a grouping for subsystems, a subsystem is the group of rivers defined by a watershead, each one of these areas will be denominated by an unique badge, for what we will have several subsystems in a group of rivers.

Horton in [4] and [5] introduced a concept of classification of streams that allows assigning whole values to streams in hydrological networks that determine their relative importance in a hierarchy of tributary bigger and smaller. An improved version of this concept was introduced later in [10]. The network is integrated by a main channel and a series of tributary whose ramifications extend from the highest parts toward the lowest parts where the glides converge [1] [9]. Fig. 8 shown a group the rivers in subsystems, with a classification of streams applying the hierarchical scheme of Horton and Strahler (HS).

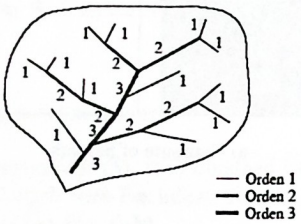


Fig.8. Classification of streams with the outline of Horton and Strahler

Once the rivers are classified in subsystems, the measurements are computed, corresponding with their sinuosity, length, order of the flowing (hierarchical class), and the range of heigths.

4 Experimental results

To obtain the measurements, we have been developed in language AML (Arc Macro Language), a set of programs to obtain the geometric characteristics. In Fig. 9, a fragment of drainage network is depicted. It is composed of lineal elements, these lineal elements possess the characteristics of any other drainage network.

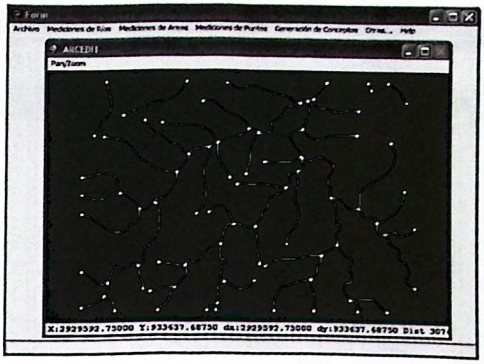
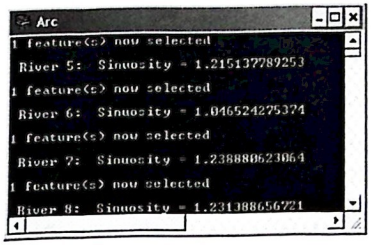
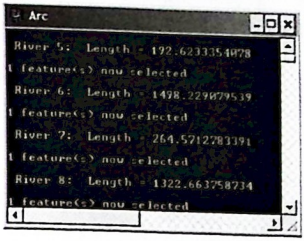


Fig. 9. Proposed drainage network

Fig. 10 (a) depicts the computation of *sinuosity measurement* for the Rivers. In addition, Fig. 10 (b) shows the compute of length measurement for the rivers.





Attributes of X2

Shape	Fnode#	Tnode#	I-altura	Condicion	Sinuosidad
PolyLine	7	4	363	Intermitente	1.004
PolyLine	8	3	240	Perenne	1.007
PolyLine	9	8	240	Perenne	1.007
PolyLine	10	5	406	Intermitente	1.025
PolyLine	12	9	240	Perenne	1.215
PolyLine	21	1	354	Intermitente	1.047
PolyLine	24	18	260	Perenne	1.239
PolyLine	25	24	260	Perenne	1.003

Attributes of X4

Shape	Fnode#	Tnode#	F-altura	I-altura	Status	Sig	Ant	Longitud
PolyLine	7	4	451	363	5	0	0	163.730
PolyLine	8	3	240	240	4	0	7	451.119
PolyLine	9	8	240	240	4	2	9	129.441
PolyLine	10	5	544	406	6	0	0	561.466
PolyLine	12	9	240	240	4	3	12	192.623
PolyLine	14	2	520	354	3	0	0	1498.229
PolyLine	18	19	260	260	4	12	15	264.571

(a) (b)

Fig. 10. (a) Sinuosity measurement, (b) Length measurement

The process of altitudes assignment, involves the use of the layer that contains the altitudes of that area, which is required to assign the heights to the nodes and later on.

This layer is used to extract the information on the altitudes. Fig. 11 depicts the computation of *heights* for the rivers.

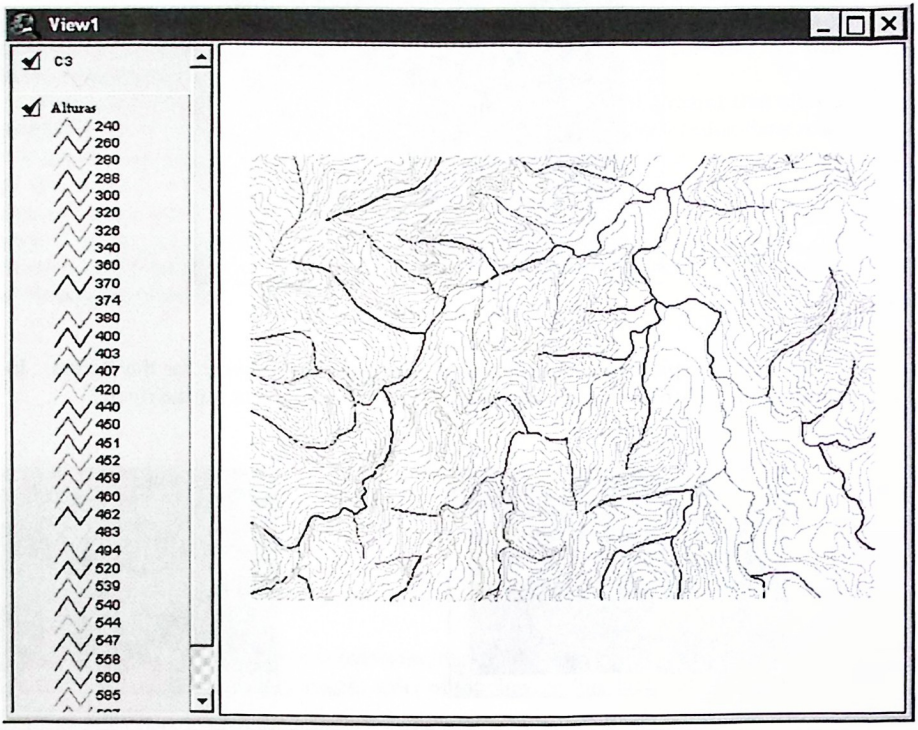


Fig. 11. Compute of heights of rivers

Fig. 12 depicts the classification of *subsystems* for the Rivers. This classification provides the measurements for subsystems and to make comparison of results later on.

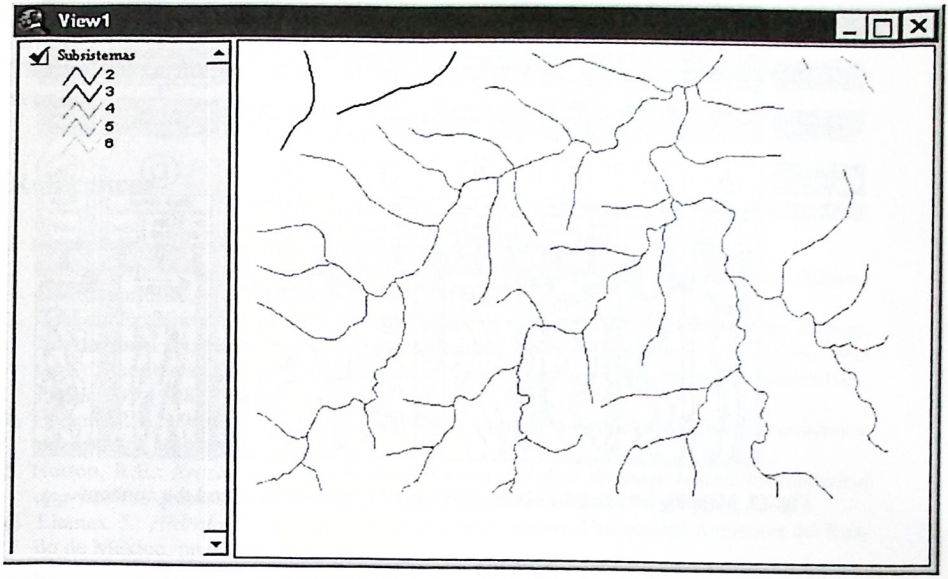


Fig. 12. Classification in subsystems

The following step is the *conceptualization*. The conceptualization process will be performed by a mapping between the range of values and the concept. In this case, the range of values represents a concept. The representative intervals are defined in the classification (for example, in the case of sinuosity a line can be *little*, *half*, *very sinuous*). The values and concepts are different according to the case study. In other words, each thematic should be represented by different intervals or concepts. Fig. 13 shows the mapping between the value of a measurement and the concept. This is defined by means of intervals (ranges).

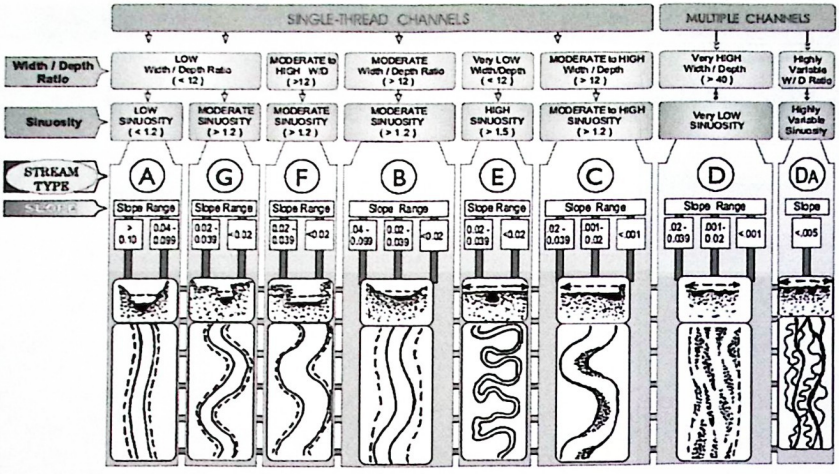


Fig. 13. Mapping between the obtained values of a measurement and the concept

It is important to mention that the generated concepts are not *ambiguous*, because they only can be applied in a specific context (geometric shape), i.e. it is different the sinuosity of highways that the sinuosity of rivers or elevation contour lines.

5 Conclusions and Future Work

In this work an approach to incorporate *semantic content* into Geographic Information Systems has been presented. The semantic data content is expressed by concepts. These concepts are generated analyzing the datasets in a specific geographic domain, and represented by concepts. The concepts represent the interpretation, and the meaning of geospatial objects. Our method is based on explicit representation of geometrical properties of geographic data by means of concepts. By using this approach, we attempt to catch or dimension *semantic contents*, which implicitly contain the spatial data and do not depend on the scale. In addition, it is possible to hierarchize the objects according to geometric characteristics.

In the future work we will measure the semantic similarity among different geospatial datasets. In our opinion, a domain conceptualization is useful to build ontologies, which represent (globally) the context of that domain, while the vocabulary of concepts and its relations describe the semantics (locally). Ontologies are very useful since they add a semantic component (the relations between different concepts), which normally is not considered in traditional GIS approaches.

Acknowledgements

The authors of this paper wish to thank the IPN, CIC, SIP and CONACYT for their support.

References

1. Black, P.E.: *Watershed hydrology.* Michigan, Ann Arbor Press, pp. 449, (1996).
2. Genesereth, M. R. and Nilsson, N.: Logical Foundation of Artificial Intelligence. Morgan Kaufmann, Los Altos, California, (1987).
3. Gruber T.: *Toward principles for the design of ontologies used for knowledge sharing*, International Journal of Human-Computer Studies, Volume 43 , Issue 5-6 Nov./Dec. 1995 table of contents, Special issue: the role of formal ontology in the information technology, Pages: 907 - 928, Year of Publication, ISSN:1071-5819, (1995).
4. Horton, R.E.: *Drainage basin characteristics*, American Geophysical Union Transactions, pp. 350-361, (1932).
5. Horton, R.E.: *Erosional development of streams and their drainage basins: hydrophysical approach to quantitative morphology.* Geol. Soc. America Bull., 56, 275-370, (1945).
6. Llamas, J.: *Hidrología general, principios y aplicaciones*, Universidad Autónoma del Estado de México, pp.626, (1989).
7. Mustière S.: What is spatial context in cartographic generalisation? *Symposium on Geospatial Theory*, Processing and Applications (2002).
8. Papadias D. and Theodoris Y.: Spatial relations, minimum bounding rectangles, and spatial data structures, *International Journal of Geographical Information Science*, vol. 11, n.2, pp.111-138 (1997).
9. Sanjuame, M.S. y Villanueva, R.J.B.: *Teoría y métodos en geografía física*, Editorial Síntesis, Madrid, pp 303, (1996).
10. Strahler, A.N.: *Quantitative analysis of watershed geomorphology.* Am. Geophys. Union Trans., 38(6), p. 913-920, (1957).
11. UNI-ZH.: AGENT Tecnical Annex [IGN, 96], Dept. of Geography, University of Zürich (1999).
12. Worboys M.F.: Metrics and topologies for geographic spatial. In Advances in GIS research II : *proc. of 7th International Symposium on Spatial Data Handling*, Kraak and Molenaar (eds), Taylor and Francis, pp.365-375, (1996).

Author Index
Índice de autores

El abrazo
1980
Jorge González Camarena
Óleo sobre tela (140 x 200 cm.)
Colección Jorge Camarena Saint-Leu

Impreso en los Talleres Gráficos
de la Dirección de Publicaciones
del Instituto Politécnico Nacional
Tresguerras 27, Centro Histórico, México, D.F.
Noviembre de 2006.
Printing 500 / Edición 500 ejemplares.

This special issue contains the papers accepted in the biannual (even years) conferences, which are satellite events of the biannual (odd years International Conferences on Geospatial Semantics (GeoS). While GeoS aim at providing a timely forum for the exchange of state-of-the art research results in the areas of modeling and processing of geospatial semantics, CONGEO focuse on the applications of geospatial semantics to next-generation spatial databases and geographic information systems (GIS) as well as personalized geospatial web services.

The Second International Conference on Geoinformatics Applications (CONGEO 2006) was held in Mexico City, November 21 - 24, 2006 at the Centro de Investigación en Computación of the Instituto Politécnico Nacional in conjunction with Magnum Congress on Computing, given an opportunity for people from the industry and academia to share their latest experiences in applications to intelligent GIS.

We hope you will enjoy the works reported in this issue and we are looking forward to see you in the next editions of CONGEO in CIC-IPN, Mexico City.

Sincerely,

Marco Moreno-Ibarra
Sergei Levachkine

ISSN: 1870-4069
www.cic.ipn.mx/rcs