Un nuovo approccio al rilevamento di errori nei geodatabase

Sandro Savino, Massimo Rumor

sandro.savino@dei.unipd.it, rumor@dei.unipd.it, Università degli Studi di Padova,
Dipartimento di Ingegneria dell'Informazione, via Gradenigo 6/B, 35131 Padova

Introduzione

Gli attuali sistemi di controllo automatico dei geodatabase si basano sulla definizione di vincoli sui dati, nella forma di vincoli di dominio, vincoli sulle geometrie, vincoli sulle relazioni tra attributi e vincoli topologici.
Questi vincoli formali permettono di rilevare errori sui domini o nei dati geometrici; non permettono tuttavia di individuare tutti gli errori di contenuto, quali ad esempio quelli di classificazione. E' possibile quindi, ed è esperienza comune, che un database formalmente corretto, cioè per il quale sono verificati tutti i vincoli di cui sopra, contenga ugualmente dati non corretti, che sfuggono ai tradizionali sistemi di validazione.
Questo lavoro illustra un nuovo approccio al rilevamento degli errori nei geodatabase che permette di rilevare quegli errori che sfuggono ai tradizionali controlli formali. Il concetto alla base di questo nuovo approccio è il concetto di "anomalia", ovvero di una informazione che, pur formalmente corretta, potrebbe essere errata.

Il concetto di anomalia e l'approccio

In un geodatabase, definiamo un'anomalia come una feature le cui caratteristiche non sono conformi ai valori tipici per quel tipo di oggetto.
Una anomalia è diversa da un errore, in quanto non viola nessun vincolo formale: un'anomalia invece è un dato che potenzialmente potrebbe essere errato. L'idea di base del nostro approccio è che cercando nei dataset dati "anomali", è possibile individuare errori potenziali impossibili da rilevare per gli strumenti tradizionali di controllo; questi dati anomali devono essere poi sottoposti ad un operatore umano che valuterà se sono effettivamente errori reali o meno.
Di seguito è fornita una classificazione delle anomalie ed alcuni esempi che rendono più chiaro il concetto di anomalia e l'approccio, fornendo anche degli spunti su come possono essere sviluppati gli algoritmi per l'identificazione delle anomalie.

Classificazione delle anomalie

La tassonomia delle anomalie individua 5 diversi tipi di anomalie, raggruppati in 3 classi. Oltre a queste, sono stati individuati altri 3 tipi di anomalie specifiche per le feature che appartengono ai network, che sono state poste in una classe a parte.

Anomalie di forma

Anomalie di regolarità: queste anomalie riguardano oggetti la cu forma non rispetta il concetto che oggetti naturali hanno forme irregolari mentre oggetti antropici hanno forme regolari; ad esempio un lago di forma rettangolare rappresenta un'anomalia.
Anomalie di dimensione semantica: queste anomalie riguardano oggetti lineari od areali la cui dimensione è troppo piccola o grande rispetto al loro "valore semantico"; ad esempio una strada lunga 1 metro o un bosco grande 10 m^2 sono delle anomalie.

Anomalie di cardinalità
Rispetto al numero totale: queste anomalie riguardano oggetti che si trovano in un numero inaspettatamente alto o basso; ad esempio trovare un grande numero di parcheggi in un'area isolata, è un'anomalia.
Rispetto alla distribuzione: per alcuni oggetti è possibile associare la tipologia ad una distribuzione caratteristica: ad esempio in un territorio urbano le chiese sono equamente distribuite mentre silos o edifici industriali sono raggruppati; nel caso in cui la distribuzione sia diversa, si rileva una anomalia.

Anomalie di posizione
Rispetto ad altre feature: queste anomalie riguardano oggetti che si trovano spazialmente troppo lontani da altri oggetti a cui sono logicamente collegati. Ad esempio un faro marittimo lontano dal mare o un rifugio di montagna in spiaggia, un aeroporto lontano da una pista di atterraggio sono tutte anomalie. Questo tipo di anomalie riguarda anche oggetti troppo vicini ad altre feature: la distanza può essere assimilata ad una dimensione semantica e quindi ricadere nell'anomalia descritta sopra. In ambo i casi la distanza va intesa non solo come misura tra due specifici oggetti ma anche come distanza da un insieme di feature che identificano un contesto semantico territoriale.

Anomalie nei network
Anomalie di classificazione: questa anomalia riguarda elementi connessi di un network nei quali si ha un cambio repentino di classificazione isolato o ripetuto.
Anomalie di connessione: partendo dal presupposto che un network dovrebbe contenere principalmente elementi connessi, rilevare molti edge isolati in un network identifica una anomalia. Quanto detto a riguardo delle anomalie di posizione vale a maggior ragione per gli elementi di un network: se la distanza tra due elementi sconnessi è troppo piccola, è una anomalia.
Anomalie di direzione: nel caso di network con edge orientati, è possibile rilevare delle anomaile analizzando le differenze nelle direzioni degli edge; ad esempio delle strade da cui si può solo uscire o un fiume che scorre in salita sono anomalie.

Conclusioni
La definizione e il rilevamento delle anomalie si basano su concetti legati alla logica e al senso comune e rappresentano un nuovo paradigma nel campo degli strumenti di controllo: invece di definire come appaiono i dati errati, è necessario definire come appaiono i dati corretti. Questo richiede un grande lavoro di definizione, ma permette di identificare errori altrimenti non rilevabili. La mole di lavoro richiesta rende l'approccio più adatto ai grandi dataset, che sono proprio quelli per i quali un controllo manuale non è fattibile.

Bibliografia
Louwsma J, Zlatanova S, Lammeren Rv, Oosterom Pv. (2006), Specifications and implementations of constraints in GIS- with Examples form a Geo-Virtual Reality System, GeoInformatica, 10:531-550.
Mas S, Reinhardt W. (2009), Categories of Geospatial and Temporal Integrity Constraints, Proceedings of the International Conference on Advanced Geographic Information Systems & Web Services, Cancun, Mexico, 146-151.
Savino S, Rumor M. (2014), Detecting errors in formally correct geodatabases, Proceedings of Eighth International Conference on Geographic Information Science, Vienna.
Udagepola KP, Xiang L, Xiaozong Y, Wijeratne AW. (2006), Review of data consistency and integrity constraints in spatial databases, Proceedings of the 5th WSEAS International Conference on Artificial Intelligence, Knowledge Engineering and Data Bases, Madrid, Spain, 348-353.