Interrogazioni in linguaggio naturale a basi dati eterogenee: l'ontologia del sistema "FuLL" nei GIS

Maurizio BOMBARA (*), Davide CALÌ (*), Ivana CALÌ (*), Emiliano GIOVANNETTI (**), Maria Vittoria MASSEROTTI (**), Chiara RENSO (**), Laura SPINSANTI (**), Giuseppe TROPEA (*)

(*) BC S.r.l. Software Company – Via Caronda n.136 – 95128 Catania
Tel. +39 0957286481 - fax +39 095444199 - info@bcsoftware.it - www.bcsoftware.it
(**) KDDLAB, ISTI CNR - Via Moruzzi 1 - 56010 Pisa
{mavie.masserotti, laura.spinanti, chiara.renso, emiliano.giovannetti@isti.cnr.it}@isti.cnr.it

RIASSUNTO

Il problema della facilità di interazione fra gli utenti e il software GIS è ampiamente conosciuto. La possibilità di utilizzare una interfaccia in linguaggio naturale verso le applicazioni GIS diviene dunque strategica. Una sfida importante, a maggior ragione nel contesto delle interfacce in linguaggio naturale, rimane quella di migrare una tale interfaccia fra basi dati eterogenee, pur se relative allo stesso dominio semantico. Si rende necessario formalizzare una struttura dati che racchiuda la conoscenza posseduta dal software sullo specifico dominio.

I risultati presentati in questo lavoro riguardano il test della tecnologia FuLL (Fuzzy Logic and Language), basata sull'approccio metodologico descritto, verso due differenti basi di dati geografiche, entrambe nel dominio della "Mobilità e Trasporti" e del "Sistema Insediativo Territoriale", tratte dai SIT delle province di Bologna e di Catania.

ABSTRACT

The interaction between users and GIS software is a known and critical issue. The importance of giving those users a natural language interface to the system is thus strategic. Moving such a NL interface between heterogeneous DBs is an even more challenging task. We accomplish this by using a domain ontology as a knowledge repository and interface between raw data and language semantics. Successful tests of FuLL's (Fuzzy Logic and Language) technology are reported, where we have used the same ontology structure and connected it to Bologna's and Catania's district geo-databases. Both databases contain (differently structured) data about the "Transportation System" and "Urban Planning" semantic domains.

KEYWORDS: *Natural Language, Data normalization, Domain Ontologies*

INTRODUZIONE

La capillare diffusione della tecnologia GIS e la nascita di una vera e propria "scienza" che ne studia sia i fondamenti teorici che i possibili campi applicativi, hanno reso questo particolare tipo di sistemi informativi un settore sul quale si concentrano notevoli sforzi investigativi. Infatti, nei GIS la componente geografica viene utilizzata come collante e rende possibile incrociare informazioni di natura diversa, generando nuova informazione.

Ma proprio la diversità di fonti dalle quali si attinge è al tempo stesso una ricchezza ed un limite. Da un lato infatti poter reperire informazione da varie sorgenti permette di ampliare la conoscenza

su un determinato dominio. Dall'altro però il reperimento delle informazioni, quando le sorgenti dati sono molteplici ed eterogenee, può essere un compito estremamente difficoltoso. Pensiamo ad esempio a problemi di formato dei dati stessi, di modello dei dati, di semantica dei termini. Tutto il problema delle basi di dati (geografiche) eterogenee è ben conosciuto e trattato ampiamente in letteratura [1].

L'interazione tra utenti (non esperti) e sistemi geografici è un'altra questione critica ben conosciuta anche in letteratura [1], [2]. Un approccio tipico per facilitare l'interazione dell'utente non esperto con un sistema informativo è l'interfaccia in linguaggio naturale. Questa, applicata in particolar modo ai GIS, permette espressioni linguistiche non precise, dichiarative, l'uso di sinonimi e di espressioni di tipo qualitativo. Al contrario, nei GIS attuali le interfacce sono di tipo quantitativo e procedurale. Ad esempio, una possibile query utente in linguaggio naturale potrebbe essere "*Dammi tutti i negozi vicini alla stazione centrale*", astraendo così da dettagli quantitativi e procedurali. Proprio quei sistemi GIS che permetteranno tali interrogazioni, consentiranno di ampliare l'audience a utenti anche non specialisti [3].

Il sistema *FuLL (Fuzzy Logic and Language*), ideato e sviluppato dalla *Software Company* BC s.r.l., con la collaborazione scientifica dei partner di progetto ed in corso di industrializzazione, realizza una interfaccia in linguaggio naturale a basi di dati geografiche. In particolare, permette l'interrogazione in linguaggio naturale a basi di dati eterogenee, posto che siano semanticamente equivalenti. La possibilità di porre query in un linguaggio integrato e uniforme permette infatti di rendere trasparenti per l'utente tutte le disomogeneità di sintassi, modelli dato, formato dei dati e terminologia delle basi di dati sorgenti. La definizione di una *ontologia* ha un ruolo centrale nell'architettura dell'intero sistema e favorisce la creazione di una interfaccia consistente e facile da utilizzare.

Per un approfondimento su come il sistema *FuLL* riesca a tradurre una query in linguaggio naturale in una Forma Logica intermedia (FL), e sia in grado di proiettare tale FL sull'ontologia ed infine a tradurla in SQL spaziale, si faccia riferimento a [4].

L'ONTOLOGIA DI *FULL*

Il componente progettato per racchiudere la conoscenza strutturata sul dominio che guida, in *FuL*L, le interrogazioni poste dagli utenti è l'ontologia. Essa si compone essenzialmente di due parti: una parte linguistica, denominata OntoLex, ed una parte di rappresentazione di dominio, denominata OntoOwl. OntoLex permette di gestire l'equivalenza semantica tra termini (sinonimia) mettendo in relazione i gruppi di sinonimi (synset) con le entità dell'ontologia che ne rappresentano il "significato" [4], mentre OntoOwl permette di rappresentare i concetti, le relazioni e gli attributi del dominio di interesse e quindi esplicita, di fatto, ciò che l'utente può chiedere.

Caratteristiche di OntoOwl e procedimento di definizione

Come abbiamo già accennato in precedenza, l'ontologia di dominio ha il compito di rappresentare tutta e sola quella parte di dominio, rappresentata nei database sorgenti, che l'utente può consultare. In altre parole fornisce una "vista", integrabile su basi di dati geografiche eterogenee, della porzione di dati che si vuole rendere consultabile.

Questa caratteristica dell'ontologia di *FuLL* deriva dalla metodologia di definizione dell'ontologia stessa. Da un lato, infatti, deve rappresentare concetti di dominio sufficientemente generali da poter essere condivisi da molteplici sorgenti dato, dall'altro deve poter dare una rappresentazione astratta di entità effettivamente presenti nei database. Deve essere, quindi, una ontologia che fonde aspetti di ontologia *applicativa* con aspetti di ontologia *fondazionale* [5].

Per ottenere questo compromesso l'ontologia è stata definita con una metodologia ibrida *top-down* e *bottom-up*. Ovvero, la definizione di concetti, relazioni e attributi è stata guidata in parte

dall'analisi dei dati sorgente (bottom-up), in parte dall'esigenza di rappresentare concetti e relazioni tipiche del dominio applicativo scelto (top-down).

Il procedimento *bottom-up* si basa essenzialmente sullo studio dei dati dei database geografici disponibili e nell'astrazione di essi da tabelle fisiche in concetti e relazioni. Questa fase è tipicamente supportata dalla presenza di un modello concettuale del database in esame. Poiché tutti i Sistemi Informativi Geografici posano su un *geodatabase*, negli ultimi decenni la modellazione concettuale, un passo molto importante nel disegnare l'architettura del database, è stata soggetta a numerosi ampliamenti per la gestione della componente spaziale.

Il formalismo classico per le basi di dati, l'entity-relationship - ER [6], è stato ampliato da Calkins [7] per riuscire a modellare efficacemente i geodatabase. Successivamente sono stati sviluppati formalismi diversi (come OMT-G, COM, MADS [4]) tutti in grado di modellare una realtà con la componente geografica, talvolta anche temporale. In particolar modo un aspetto centrale è quello della topologia. Oggi, nel tentativo di proporre degli standard, alcuni software GIS utilizzano il linguaggio di modellazione UML (Unified Modeling Language) per consentire il dialogo tra geodatabase diversi [8].

Nell'ambito del sistema *FuLL*, per la sua semplicità e per la facilità di tradurre i concetti espressi in un'ontologia, è stato usato il formalismo ER ampliato da Calkins. Questo formalismo coniuga la semplicità d'uso con l'ampiezza dell'informazione e la gestione della loro *georeferenziazione*.

Il ruolo della modellazione concettuale in *FuLL*, utilizzando tecniche di *reverse engineering* appositamente studiate, è stato quello di costruire un modello partendo da banche dati di natura diversa, costruite senza una vera e propria architettura. Il modello concettuale e logico della banca dati presa in esame supporta la fase *bottom-up* della definizione dell'ontologia di dominio.

Nell'approccio *top-down* si procede invece ad un'analisi ad alto livello del dominio investigato. In particolare, in questa fase si prendono in esame contemporaneamente diversi aspetti: viene effettuato uno studio su documenti descrittivi del dominio per estrapolarne i concetti fondamentali; vengono estrapolate le conoscenze di esperti sul dominio; viene individuato un glossario di termini significativi del dominio per rappresentare i concetti chiave e per fornirne una interpretazione univoca, aggregando eventualmente alcuni concetti di "base" individuati nella fase di costruzione bottom-up in concetti di più alto livello, come è stato fatto, ad esempio, nella fusione dei concetti di "elemento stradale" e "giunzione stradale" nell'unico concetto di "strada". Infine si integrano gli elementi individuati dalle due metodologie in un'unica ontologia di dominio.

L'ARCHITETTURA A LIVELLI

Altro elemento progettuale importante del sistema *FuLL* è stato quello concernente il collegamento dell'ontologia alla base di dati dalla quale i dati devono in ultima analisi essere estratti. Tale collegamento è stato strutturato in una architettura a più livelli, partendo da un livello più astratto sino ad uno schema logico a più basso livello, a ridosso dei dati stessi del database. Lo scopo è quello di poter sostituire la base di dati sottostante, a parità di dominio, senza bisogno di riscrivere l'ontologia. Ad ogni entità che appare nell'ontologia, sia essa una classe, una relazione o un attributo) deve corrispondere un collegamento con un elemento di quello che abbiamo definito *schema logico del DB target*, realizzato normalizzando, quando necessario, alcuni dati del database (DB) stesso. Vi sono dunque due distinti livelli di astrazione, presenti in *FuLL*, per collegare l'ontologia ai dati: un collegamento (realizzato tramite *view* SQL) dell'ontologia allo schema logico del DB e lo schema logico del DB stesso. Questa architettura serve per disaccoppiare l'ontologia dallo specifico DB, permettendo al sistema di restare quanto più possibile generale. Così facendo possiamo sostituire i dati provenienti da un altro DB (posto che presenti dati appartenenti allo stesso specifico dominio, ad esempio strade, imprese, agriturismi, etc...), cambiando di conseguenza la costruzione delle view di collegamento e realizzandone opportunamente lo schema logico.

La realizzazione dello schema logico del DB dipende dallo specifico DB target e prevede uno studio dei dati per normalizzarli, ove necessario. Potrebbe infatti non esserci una diretta coincidenza fra i termini ed i concetti del dominio, con la rappresentazione presente nel database di tali concetti e relazioni. Occorre cioè normalizzare il DB introducendo eventuali nuove tabelle, corrispondenti a classi dell'ontologia per le quali non esisteva una tabella di riferimento. Nei casi più complessi può accadere anche che l'informazione sia organizzata nel DB in maniera così differente rispetto all'ontologia da dover mettere a punto procedure per estrarre i dati di interesse da una o più colonne specifiche, e trasferirle in righe distinte, allo scopo di riorganizzarli in altre tabelle maggiormente rispondenti alla struttura ontologica, costruendo così lo schema logico opportuno.

Una volta effettuata la normalizzazione dei dati nel corrispondente schema logico, è necessario collegare le entità dell'ontologia con i dati presenti nello schema logico del DB. Tale collegamento viene realizzato tramite dei comandi SQL di create view che seguono uno schema standard. Le view sono realizzate in maniera semi-automatica dal software, creando i necessari adattamenti e *cast* fra i tipi di dato presenti nel DB.

L'INTERFACCIA IN LINGUAGGIO NATURALE DI *FULL*

Il problema fondamentale nell'utilizzo delle applicazioni GIS è che diversi utenti hanno finalità di consultazione diverse. Si pensi ad esempio ai diversi impieghi dello stesso GIS che potrebbero fare dirigenti, tecnici (medici, volontari, spesso non informatici) in situazioni di emergenze di protezione civile ed anche, d'altra parte, cittadini e imprese nella fruizione del patrimonio informativo territoriale gestito dalle pubbliche amministrazioni. Spesso si tratta di utenti con poca dimestichezza verso i meccanismi tipici delle interfacce dei prodotti software. Inoltre nei GIS più complessi, per ottenere le informazioni che si vogliono, bisogna scrivere complesse query SQL oppure usare una interfaccia visuale, usufruendo solo delle query che il progettista del software ha pensato di mettere a disposizione. I vantaggi di una interfaccia in linguaggio naturale appaiono dunque immediati.

Confronto fra interfaccia in linguaggio naturale e interfaccia tradizionale

Vediamo ora un esempio di esecuzione di una query in linguaggio naturale col sistema *FuLL* a confronto con un sistema GIS (in questo caso è stato preso ad esempio ArcGIS di ESRI, ma la procedura è generale per ogni sistema geografico).

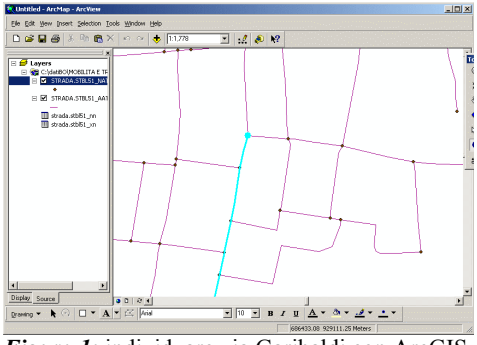

Figura 1: individuare via Garibaldi con ArcGIS

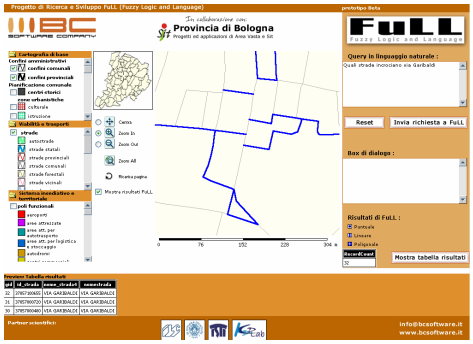

Figura 2: quali strade incrociano via Garibaldi?

Supponiamo che il dominio considerato sia "mobilità e trasporti" e "sistema insediativi territoriale", e che la base di dati contenga informazioni sul reticolo stradale. La query potrebbe essere: "*Quali strade incrociano Via Garibaldi*"?.

La sequenza di operazioni da effettuare con il sistema di interfaccia tradizionale (rappresentata in Figura 1) è tipicamente lunga e articolata:
1) Occorre individuare l'identificatore della strada che ha "via Garibaldi" come toponimo 2) Si deve accedere alla tabella toponimi per avere l'identificatore della strada 3) Occorre identificare gli archi (elementi stradali) che la compongono accedendo alla tabella archi 4) Una volta selezionati tutti gli archi relativi a via Garibaldi possiamo trovare le giunzioni relative (cioè gli incroci) dalla tabella giunzioni 5) Occorre ripetere all'indietro tutte questa procedure per recuperare il nome strada degli archi relativi alle giunzioni trovate.

È chiaro come una procedura di questo tipo risulti estremamente difficoltosa per un utente non esperto. Sarà inoltre necessario automatizzare questa sequenza di operazioni codificandole in una procedura.

Come mostrato in *Figura 2*, l'approccio di *FuLL* invece permette di rendere completamente trasparente per l'utente la complessità procedurale della risoluzione di ogni singola query. In questo caso, infatti, la query potrà essere risolta rappresentando in OntoOwl la relazione *incrocia* del concetto *Strada*, e stabilire un collegamento tra tale relazione e una view del livello logico del database che rappresenti gli incroci delle strade.

RISULTATI

I due domini presi in esame sono stati quelli relativi a "mobilità e trasporti" e "sistema insediativo territoriale". Essi sono risultati domini particolarmente complessi, in termini di relazioni fra le entità coinvolte.

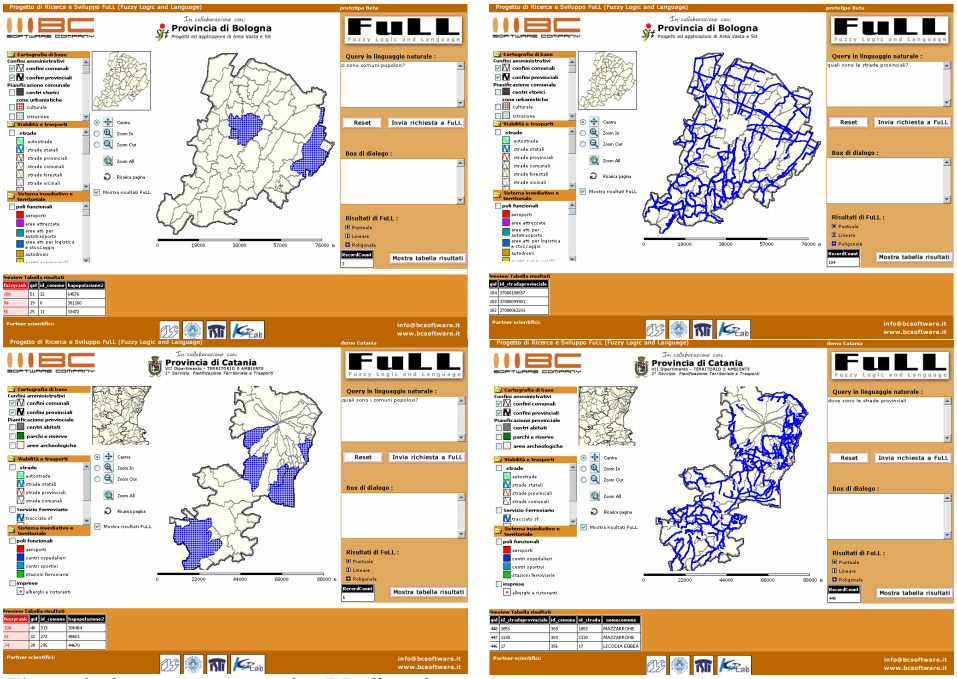

Figura 3: le stesse query su due DB diversi

E' stata realizzata un'unica ontologia per entrambi i domini, che è stata collegata con successo a due distinti database geografici estratti dal SIT delle province di Bologna e Catania. Nelle

interrogazioni d'esempio riportate in *Figura 3* ("*dove sono le strade provinciali*" e "*quali sono i comuni popolosi*") è possibile notare come l'interfaccia in linguaggio naturale sia stata collegata alle due basi di dati distinte, offrendo la possibilità di interrogarle in maniera omogenea. Notiamo che i risultati ottenuti sono di facile interpretazione su entrambi i domini nonostante la differenze in termini di modello dei dati e formato fossero estremamente diversificati nei due SIT.

Si noti come *FuLL* è in grado di distinguere fra espressioni del tipo "*quali sono*" rispetto a "*dove sono*", fornendo nell'ultimo caso anche informazioni circa la localizzazione delle entità, così come il fatto di poter utilizzare diverse forme ("*ci sono*", "*quali sono*", …) per comunicare la richiesta al sistema.

CONCLUSIONI E SVILUPPI FUTURI

L'approccio metodologico qui presentato è in corso di valutazione accurata tramite uno studio sistematico su una gamma più ampia e diversificata di geo-database, estesi ad altri domini, come quello della Protezione Civile. Al contempo viene valutata l'efficienza dell'interfaccia naturale rispetto ad approcci più tradizionali per interrogare i DB. L'introduzione di concetti temporali, e l'interrogazione di basi dati spazio-temporali rappresenta il naturale sviluppo della tecnologia *FuLL*, così come l'integrazione con strumenti di fruizione più pervasivi quali telefoni cellulari (via SMS/MMS) e dispositivi mobili avanzati.

RINGRAZIAMENTI

Si ringraziano tra i partner scientifici che hanno collaborato al progetto *FuLL*:

- C.N.R., Istituto di Linguistica Computazionale (I.L.C.)
- Università di Pisa, Dipartimento di Linguistica

e le pubbliche amministrazione che hanno reso disponibili i dati per la sperimentazione condotta

- Provincia di Bologna, Progetti ed Applicazioni di Area Vasta e SIT
- Provincia di Catania, 1° Servizio Pianificazione Territoriale e Trasporti del VII Dipartimento

BIBLIOGRAFIA

Sheth, A P., 1999. "Changing Focus on Interoperability in Information Systems: From System, Syntax, Structure to Semantics", Edited by Goodchild, M. F. Egenhofer, M. J., Fegeas, R., and Kottman, C. A., *Interoperating Geographic Information System*s, Kluwer

Medyckyj-Scott, D. & Hearnshaw, H., 1993. "Human Factors in Geographical Information Sys tems", *Belhaven Press*, London

Max J. Egenhofer, A. & Rashid, B. & Shariff, M., 1998. "Metric details for natural-language spatial relations", *ACM Transactions on Information Systems (TOIS)* Volume 16, Issue 4, pp. 295 – 321

Bartolini R., Caracciolo C., Giovannetti E., Lenci A., Marchi S., Pirrelli V., Renso C., Spinsanti L., 2006. "Creation and Use of Lexicons and Ontologies for Natural Language Interface to Databases", *in proceedings of 5° conference on Language Resources and Evaluation (LREC 2006)*

Guarino N., 1997. "Understanding, building, and using ontologies: A commentary to "Using Ex plicit Ontologies in KBS Development", by van Heijst, Schreiber, and Wielinga." *Inter national Journal of Human and Computer Studies* 46: 293-310

P.P Chen., March 1976. "The entity-relationship model - toward a unified view of data", *ACM Transactions on Database Systems*, vol. 1, no. 1, pp. 9-36

Hugh W.Calkins, January 1996. "Entity Relationship Modeling of Spatial Data for Geographic Information Systems", *International Journal of Geographical Information Systems*

Zeiler, Michael, 1999. "Modeling Our World – The ESRI Guide to Geodatabase Design", *ESRI Press*