

Graph Databrowsing

Sistemi di analisi e accesso ai dati basati sui grafi

di Ernesto Lastres*, Stefano Corsi

*Sister - Sistemi Territoriali

Until today, those intending to make an analysis of combinations of data could use two approaches: "Data Retrieval", the objective exploration of the data, or "Data Mining", a knowledge-mining process using the application of algorithms to identify "hidden" associations. CINECA, in collaboration with its partner Sistemi Territoriali (Sister), proposes a third approach, called Graph Databrowsing, which permits highlighting of "hidden" relations between data through graphic representation with a simple and effective interface. This means of exploring information and, in particular, the interrelations between the entities under investigation (persons, property, assets, etc.) is finding interesting application in the fight against fraudulent conduct, both criminal and fiscal/administrative (taxes, insurance). This article illustrates the functional principles underlying Graph Databrowsing, including practical examples intended to explain the area of applicability of this information investigation system.

Fino ad oggi chi intende effettuare delle analisi su un insieme di dati dispone di due approcci: il "Data Retrieval", ovvero si cercano conferme attraverso l'esplorazione oggettiva dei dati, o il "Data Mining", un processo di estrazione di conoscenza tramite l'applicazione di algoritmi che individuano le associazioni "nascoste" tra i dati.

Il Dipartimento di Business Intelligence di CINECA, in collaborazione con il partner Sistemi Territoriali (Sister) propone un terzo approccio, denominato Graph Databrowsing, che consente di evidenziare le relazioni "nascoste" tra i dati mediante una rappresentazione grafica caratterizzata da un'interfaccia semplice ed efficace.

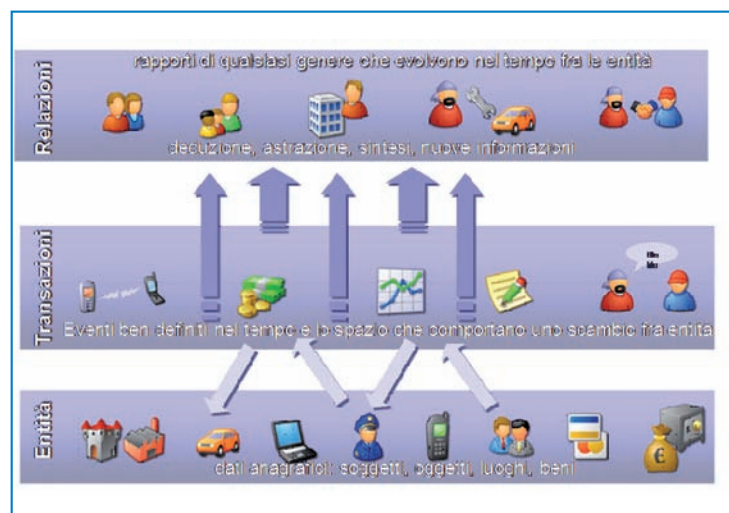
Questa modalità di esplorazione delle informazioni e, in particolare, delle interrelazioni tra le entità oggetto di indagine (persone, proprietà, beni, ecc.) sta trovando interessanti applicazioni nella lotta ai comportamenti fraudolenti, siano essi di natura criminale che di tipi fisca-

le/amministrativo (fisco, assicurazioni).

Il Graph Databrowsing è una tecnologia che fonde insieme tecniche di navigazione ed esplorazione di dati multidimensionali con concetti della teoria dei grafi, per scoprire collegamenti impliciti fra dati logicamente scon-

Graph Databrowsing
Data analysis and access systems based on graphs

Schema architetturale di base del sistema



nessi. Questo innovativo approccio permette di costruire i legami (o relazioni) che sussistono tra diverse entità contenute in un database, consentendo ad un operatore di individuare delle eventuali criticità o evidenze nascoste in tali relazioni. La tecnologia è orientata a risolvere problemi di analisi di dati che presentano un'elevata complessità intrinseca dovuta alla presenza di un gran numero di relazioni fra entità. In particolare si rende necessaria quando fra le entità esiste un'alta cardinalità e dinamicità nei legami, i quali possono anche variare nel tempo.

Con l'ausilio di un'interfaccia grafica avanzata basata sui grafi, si può ottenere un accesso immediato ai dati, nonché rappresentare l'informazione in modo da evidenziare legami nascosti fra i dati. L'alta dinamicità e flessibilità di questo tipo di interfaccia rende le applicazioni su cui si basa, uno strumento potente di analisi e di indagine.

Applicazioni così concepite costituiscono uno strumento di analisi avanzato di dati anagrafici, di business e di relazioni tra soggetti e/o entità su un database centralizzato suddiviso logicamente per indagini, dove gli operatori collaborano per portare a termine i propri lavori investigativi.

Che limitazioni possono avere i modelli relazionali puri?

- Non prevedono strumenti di analisi di lunghe catene di relazioni poiché sono basati nell'algebra relazionale la quale non permette in modo diretto di eseguire interro-

gazioni che seguano lunghe catene di relazioni ricorsive e dinamiche. Implementazioni del genere su DBMS relazionali sono altamente inefficienti e costose.

- Non dispongono di metodi di interazione che favoriscano l'accesso ai dati con complesse interrelazioni. Ovvero non possiedono strumenti e metodi di accesso ai dati che ne facilitino l'accesso e l'esplorazione.
- Non dispongono di algoritmi ottimali per ricercare catene di relazioni. Se si ragiona sulla base dell'algebra relazionale (Joins, Unions, Products, ecc.) risulta impossibile risolvere problemi di natura molto complessa.
- Non dispongono di un controllo semantico sulle strutture dati. Per poter garantire il controllo semantico, oltre ad una struttura dati molto flessibile c'è bisogno di ragionatori e strumenti di inferenza che permettano la validazione semantica dei dati e la derivazione di conoscenza a partire dall'informazione di base, e questo i DBMS tradizionali non lo garantiscono.

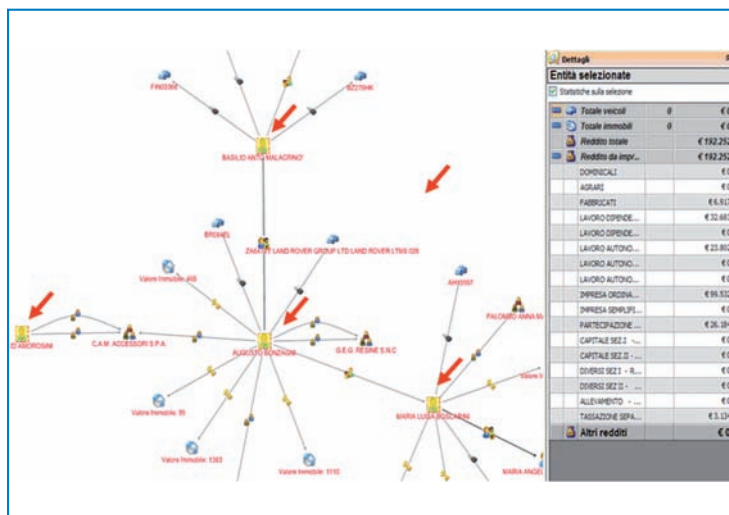
La soluzione

Per risolvere i problemi su esposti possiamo percorrere le seguenti strade:

Teoria dei grafi: percorsi, identificazione di collegamenti. L'informatica e la teoria dei grafi ci danno gli strumenti giusti per immagazzinare e gestire in modo ottimale strutture dati molto complesse con un'alta cardinalità e variabilità nelle relazioni. Inoltre fornisce la teoria e l'implementazione di algoritmi sofisticati e ottimizzati per la risoluzione di problemi quali: identificazione di collegamenti, percorsi, analogie, raggiungibilità, ecc.

Grafi interattivi: rappresentazione e manipolazioni di informazione, accessibilità. Le interfacce grafiche interattive a forma di grafo permettono di rappresentare l'informazione in modo naturale e immediato, ottenendo a colpo d'occhio non solo informazione dei dati di interesse, ma anche del loro contesto. Questo modo di navigare l'informazione, oltre ad essere più immediata e attraente, risulta più libera da schemi, ordini e vincoli di sistema, permettendo agli utilizzatori di fare una navigazione multidimensionale totalmente libera

Una rappresentazione, mediante grafi, di entità e relazioni



in base alle proprie esigenze. In questo modo si potrebbero scoprire velocemente collegamenti impliciti fra gli oggetti che altrimenti sfuggirebbero all'attenzione dell'operatore.

Ontologie: analisi semantica, valutazione e controllo dell'informazione. Una ontologia è la rappresentazione di un insieme di concetti in un dominio ben definito e delle relazioni fra di essi. Le ontologie possono aiutarci in questo modo a definire il significato dei dati che trattiamo e di conseguenza ci permettono il ragionamento automatizzato su di essi. Le ontologie sono strettamente legate al concetto di grafo, sia in termini di definizione dei concetti e relazioni che in termini del trattamento dell'informazione.

Il Modello

Il modello del Graph Databrowsing è suddiviso in tre livelli strettamente collegati:

Livello Entità

È costituito da tutti gli Attori materiali o non, che fanno parte del mondo che vogliamo modellare. Un attore/entità è un esecutore di azioni e di solito ha un ruolo all'interno del nostro modello. Esempi di attori possono essere: una persona fisica o giuridica, un telefono cellulare, una carta di credito, un conto in banca, una compagnia assicuratrice, una polizza, ecc. Le entità hanno uno stato che può variare nel tempo e quindi possono avere una storia, e possono essere strutture dati complesse con relazioni statiche fra di loro (ad esempio: Persona ↔ Carta di Identità).

Livello Transazionale

È costituito da eventi ben definiti nel tempo e lo spazio e che coinvolgono una o più entità (attori) del nostro modello. Questi eventi spesso comportano lo scambio di informazione fra entità (da cui evento transazionale). Possono essere eventi: una telefonata, un bonifico, un pagamento con Carta di Credito, un acquisto, un sinistro, un furto, un attentato, ecc.

Livello Relazionale

Questo è un livello di sintesi che permette di modellare i collegamenti di diverso tipo che esistono fra le entità (attori) del nostro modello. Per questo motivo viene visto come una sintesi di quello che accade nel mondo reale

(sintesi del livello transazionale). Queste relazioni possono essere derivate a partire dagli attributi degli attori stessi e dalle informazioni registrate nel livello transazionale. Diversamente dalle transazioni, le relazioni non sono ben definite nel tempo e nello spazio, semplicemente possono avere un intervallo temporale di validità. Le relazioni possono essere informazioni di prima mano oppure possono essere derivate a partire dalla informazione di base (entità e transazioni). Esempi di relazioni possono essere: è figlio di, è amico di, è proprietario di, è usato da, è socio di, è conoscente di, è coniuge di, è capo di, è impiegato di, ecc.

Il Grafo

In quest'ottica, il grafo è costituito da entità come nodi del grafo, e da relazioni come archi di esso. Gli eventi e transazioni sono dettagli riferiti agli archi del grafo.

L'implementazione

Astrazione

Per implementare il nostro modello è necessario modellare secondo questa filosofia (entità – eventi – relazioni), astraendo e semplificando gli attributi per concentrarci sui problemi da risolvere. Nel fare ciò buona parte dei dati di business vengo esclusi o semplificati.

ETL (Extraction Transformation and Loading)

Per questo motivo è necessario un meccanismo di estrazione di informazione, trasformazione e caricamento che prenda i dati di business di interesse, li trasformi e semplifichi nel modo adeguato e poi li carichi nel nostro modello. L'estrazione dovrebbe avvenire in automatico a partire dai Datawarehouse di business e da informazioni aggiuntiva fornita in supporti vari e con mezzi vari.

Derivazione

Una volta popolato il nostro modello possiamo eseguire processi di derivazione automatici più inerenti al nostro modello di Graph Databrowsing che permettono di determinare relazioni implicite nei dati.

Lo strumento di analisi interattiva

Permette di visualizzare il grafo (entità e relazioni) in modalità grafica, con la possibilità di

espandere o collassare il grafo e navigare le relazioni, nonché calcolare percorsi e collegamenti fra entità (calcolo di intermediari o di raggiungibilità). In ogni momento è possibile sapere i dettagli di relazioni ed entità esposte. Questo modo di accedere all'informazione viene comunemente conosciuta come link o network analysis. Lo strumento si arricchisce con la possibilità di fare ricerche alfanumeriche e la possibilità di impostare filtri su entità e relazioni.

Quando allo strumento di analisi diamo la possibilità di rappresentazione sintetica o grafica di variabili numeriche collegate a entità e relazioni, allora lo strumento diventa un

potente mezzo di analisi multidimensionale dove possiamo ottenere statistiche di sintesi in un numero praticamente illimitato di modi. In questa maniera, alla conoscenza di fatti e legami, aggiungiamo la certezza dei calcoli statistici (nella figura in basso, un esempio di statistica di sintesi del reddito per tipologia sui soggetti selezionati nel grafo).

Per ulteriori informazioni:
info.bi@cineca.it

doi:10.1388/notizie-61-08

Un caso pratico: esempio di analisi per frode assicurativa

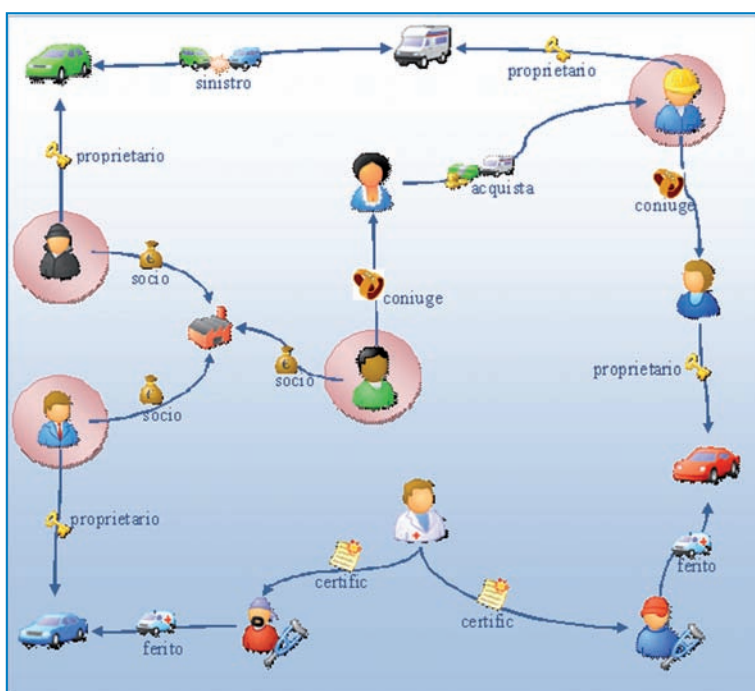
Definizione del problema. Nel mondo le principali aree dove si concentrano le frode assicurative riguardano le assicurazioni RC Auto, le assicurazioni sanitarie e di proprietà. Consideriamo il caso delle assicurazioni RC Auto. Nel mondo esistono diversi modus operandi per questi tipi di frodi, che vanno dalla dichiarazione di incidenti falsi e danni inesistenti alla falsificazione di documenti e inflazione di fatture di riparazione. Tutto questo con l'obiettivo di ottenere un risarcimento non dovuto.

È importante segnalare che se questi eventi fraudolenti capitano sporadicamente, senza precedenti ed evidenti indizi di rischio, nel peggiore dei casi passano inavvertiti. In ogni modo, dietro ogni attività criminale c'è un pattern di comportamento, che si ripete nel tempo e che può essere individuato facendo un'analisi dei soggetti, le loro relazioni e degli eventi (anche indiretti) che li coinvolgono. In questo modo è possibile determinare la probabilità che un singolo sinistro sospetto sia parte di una più ampia attività di frode.

Gli attori. Le entità coinvolte in questo tipo di analisi possono essere distinte su due livelli, dove nel primo troviamo: la polizza (il mezzo assicurato) - Il contraente - I veicoli coinvolti nel sinistro - Tutte le persone che hanno subito un danno - Tutti i beni materiali che hanno subito un danno e i loro proprietari. Nel secondo livello possono essere coinvolti: Polizze (mezzi) di persone fisiche o giuridiche coinvolte - Familiari delle persone coinvolte - Soci in affari - Aziende (datori di lavoro) - Aziende (proprietà) - Medici (che rilasciano certificati) - Ispettori -

Gli eventi. Gli eventi sono: in primo luogo il sinistro in questione ed eventuali sinistri relazionati alle persone coinvolte in un primo o secondo livello; poi abbiamo come eventi correlati gli atti di compravendita e dichiarazioni e/o denunce presso le forze dell'ordine.

Le relazioni. Sono le relazioni di partecipazione a sinistro, di assicurazione, di proprietà, di parentela, ecc.



Rappresentazione grafica del caso: gli elementi cerchiati sono in associazione per truffare le assicurazioni e possono contare sulla complicità di medici, familiari ed amici