

Informatica per Statistica

Riassunto della lezione del 28/11/2012

Igor Melatti

Introduzione alla progettazione concettuale di basi di dati

- Questo riassunto è da intendersi come un commento alle slide BD2002-06.PDF
 - non tutte le slide sono commentate
 - quelle che possono essere saltate sono esplicitamente indicate
- Slides 2-3: per realizzare un sistema informativo (non quindi necessariamente informatico) non si può andare a caso: occorre una fase di *progettazione*
 - progettare vuol dire: capire cosa bisogna fare (studio di fattibilità ed analisi dei requisiti, talvolta chiamate *specifiche*), capire che strada si può seguire per realizzarlo (progettazione vera e propria), e poi realizzarlo effettivamente (realizzazione e collaudo); dopodiché, lo si può mandare in esecuzione (funzionamento)
 - non necessariamente va tutto liscio: può succedere che in fase di realizzazione vengano scoperte nuove necessità, e che quindi occorra aggiornare la fase di analisi dei requisiti, più ovviamente tutte le fasi che seguono
 - una volta realizzato il sistema, il lavoro non è finito: occorre effettuare la manutenzione, ovvero monitorare il sistema per rilevare guasti
 - in questo corso ci si occupa quasi esclusivamente della progettazione, che nel caso delle basi di dati è la *progettazione dei dati*
- Slide 9: in questo corso si parlerà delle progettazioni concettuale e logica, quindi di come passare dai requisiti della base di dati (una descrizione informale delle informazioni che dovranno essere rappresentati dalla base di dati, ovvero la specifica) al modello logico, passando per quello concettuale
 - non verrà trattato il modello fisico

- Slide 11: ogni livello di progettazione dei dati ha il suo modello dei dati
 - ovvero, un insieme di costrutti che possono essere usati per descrivere i dati
 - nel caso della progettazione concettuale, si usa un modello concettuale dei dati; in questo corso verrà usato il *modello entità – relazione* (*entity–relationship*, di gran lunga il più usato a questo livello)
 - * qui i costrutti sono quelli di entità, relazione e attributo, collegabili tra loro in modi che verranno trattati più avanti
 - nel caso della progettazione logica, si usa un modello logico dei dati; in questo corso verrà usato il *modello relazionale* (volendo ce ne sono altri, come il modello gerarchico e quello ad oggetti)
 - * qui i costrutti principali sono le relazioni (o tabelle) intestate
 - nel caso della progettazione fisica, si usa un modello fisico dei dati; in questo corso non verrà discusso alcun modello fisico (che è interno ai DBMS, e quindi trasparente per il progettista di basi di dati)
- Slide 12: differenza generale tra schema di base di dati (ne descrive l'organizzazione) e istanza di base di dati (i dati veri e propri)
- Slide 16: l'utente, di un DBMS, vede solo lo schema logico (ovvero le tabelle), mentre il DBMS stesso si occupa della parte fisica
 - lo schema concettuale qui non si vede perché è servito, in precedenza, a costruire lo schema logico
- Slide 17: come già detto in precedenza, la progettazione concettuale parte da descrizioni verbali (in linguaggio comune) e genera uno schema E/R, dal quale poi la progettazione logica genera un insieme di tabelle intestate (con annotazioni, come si vedrà)
- Slides 20-22: i principali costrutti di uno schema entità -relazione (E/R in breve)
 - un'entità è un concetto in qualche modo autonomo, che può cioè esistere in modo indipendente da altre entità
 - una relazione esiste solo per collegare due (o anche più) entità, per esprimere un loro collegamento
 - * talvolta la relazione del modello concettuale dei dati viene chiamata anche *associazione*, per distinguerla dall'omonima (e diversa!) relazione del modello logico dei dati
- Slide 24: le entità dello schema E/R rappresentano concetti astratti, non dati concreti (che possono essere visti come *istanze* di un'entità)

- diverso dall'istanza di un database nel modello logico: qui ogni elemento concettualmente appartenente all'entità è un'istanza a sé stante; nel modello logico, invece, un'istanza è formata da *tutte* le righe contenute in una tabella
- Slide 26: anche se potrebbe venire naturale usare il plurale, per i nomi delle entità si usa il singolare
 - l'idea è di dire “l'entità Impiegato”, non “l'entità degli impiegati”: chiarisce meglio che si tratta di un'astrazione
- Slides 28-29: anche se potrebbe venire naturale usare verbi, per i nomi delle relazioni si usano sostantivi
 - l'idea è di dire “la relazione Residenza”, non “la relazione risiede in”
- Slides 30-31: per le relazioni è importante definire le occorrenze previste
 - si tratta di una relazione simile a quelle matematiche, quindi se una relazione R associa due entità A e B , può succedere che ad un'istanza di A sia correlata a nessuna, una o molte istanze di B , e viceversa
 - si può facilmente generalizzare al caso in cui la relazione colleghi più di 2 entità
 - occorrerà specificare, per entrambi i versi di una relazione, in quali casi (uno, nessuno o molti) si prevede che ci si trovi (vedere più avanti)
 - in ogni caso, un'istanza di una relazione è una n -upla di istanze da ciascuna delle entità coinvolte
 - nella maggior parte dei casi una relazione collega 2 entità, quindi $n = 2$, n -upla vuol dire coppia e un'istanza della relazione è una coppia di istanze, una da ciascuna entità
- Slide 32: il secondo schema E/R (paziente - visita - medico) potrebbe non essere corretto se si vuole rappresentare una realtà in cui un medico visita più volte lo stesso paziente
 - se questo è vero, infatti, allora avremmo una stessa istanza di medico collegata più volte alla stessa istanza di paziente, e questo non si può fare
 - i collegamenti uguali collassano in un collegamento solo, con perdita quindi di informazione
 - il che vuol dire violare le specifiche, cioè l'analisi dei requisiti da cui si parte
 - non è troppo difficile correggere questa situazione, basterà aggiungere un attributo alla relazione (vedere più avanti)

- Slide 38: un attributo è un concetto che contiene un dato sostanzialmente singolo, mentre un'entità contiene dati solitamente composti
 - deve obbligatoriamente essere associato ad un'entità o ad una relazione
 - ogni attributo ha un *dominio*, ovvero un insieme da cui può prendere valori (stringa, numeri interi, numeri frazionari...)
 - nella progettazione concettuale, il dominio resta sottinteso, ovvero non viene specificato nel diagramma E/R
 - un attributo messo su una relazione equivale a mettere un'etichetta a ciascuna n -upla di istanze correlate
 - * diventa possibile avere n -uple ripetute
 - * per esempio, se alla slide 32 si aggiunge un attributo “data” alla relazione visita, allora è possibile rappresentare correttamente il caso in cui un medico visita più volte lo stesso paziente: basta che lo faccia in date diverse
 - * una stessa coppia, con etichette diverse, diventa due coppie diverse
 - * ad esempio l'insieme di istanze {(MedicoA, PazienteA), (MedicoA, PazienteA)} non va bene, ma l'insieme {(MedicoA, PazienteA):01/01/2010, (MedicoA, PazienteA):01/06/2010} sì
- Slide 44: su ogni “lato” di una relazione occorre specificare due numeri:
 - cardinalità minima m : vuol dire che ogni istanza dell'entità che si trova su quel lato è presente in almeno m istanze della relazione (quindi è collegata almeno m volte con qualche istanza delle altre entità)
 - cardinalità massima M : vuol dire che ogni istanza dell'entità che si trova su quel lato è presente in al più M istanze della relazione (quindi è collegata al più M volte con qualche istanza delle altre entità)
- Slide 49: la semplificazione di slide 46 fa sì che si possano classificare le relazioni (binarie, ovvero tra 2 entità, ma si può generalizzare) in queste 3 categorie
- Slide 54: cardinalità per gli attributi, sempre minima e massima
 - se la minima è zero, vuol dire che per quell'attributo non sempre si sarà in grado di fornire un valore (vedere il discorso sul valore NULL nella lezione che seguono)
 - se la massima è superiore ad 1, vuol dire che l'attributo è multivalore
- Slide 56: per ogni entità A , occorre definire un *identificatore*

- ovvero, un sottoinsieme degli attributi di A (*identificatore interno*) e/o una relazione in cui A sia coinvolta con cardinalità $(1, 1)$ (*identificatore esterno*) che *si prevede* possano distinguere ogni istanza di A dalle altre istanze di A
 - si consideri un'entità Aula con attributi Nome, Capienza, ProiettorePresente, Area, e si supponga che sia collegata con un'altra entità Edificio da una relazione Collocazione
 - ovviamente, le cardinalità di Collocazione saranno $(1, 1)$ dalla parte di Aula (ogni aula si trova esattamente in un edificio) e $(1, N)$ dalla parte di Edificio (ogni edificio ha almeno un'aula, e in generale ne avrà molte)
 - la capienza, il fatto che il proiettore sia presente o meno, o l'area non sono sufficienti ad individuare un'aula: è molto facile pensare ad aule diverse in cui tali informazioni coincidano (anche tutte contemporaneamente); pertanto, i rispettivi attributi non possono essere usati come identificatori
 - il nome è sicuramente unico all'interno dello stesso edificio, ma può non esserlo all'interno della struttura nel suo complesso; cioè, possono esserci due edifici diversi che hanno due aule (ovviamente diverse!) con lo stesso nome; siccome in Aula non si distingue tra questi due casi, è necessario mettere l'identificatore esterno Nome + Collocazione
- Slide 61: la generalizzazione è una relazione “sbilanciata” tra un'entità E e altre entità (anche una sola), ed indica che E comprende le altre entità come “casi particolari”
 - Slide 65: se un'entità E ha degli attributi, ed è specializzata in altre entità, allora anche queste ultime hanno gli attributi di E (li “ereditano”)
 - lo stesso vale per altri concetti
 - nello schema E/R, gli i concetti ereditati vengono rappresentati solo sull'entità E , ed è sottinteso che ne godono anche le specializzazioni
 - Slide 66: esempi
 - una generalizzazione di Università in Statale e Privata è totale (non esistono università né statali né private) e esclusiva (ogni università è o statale o privata)
 - una generalizzazione di Università in Settentrionali e Meridionali non è totale (esclude quelle del centro) e esclusiva (ogni università è o settentrionale o meridionale)
 - una generalizzazione di Università in MedioPiccole e MedioGrandi è totale e non esclusiva

- una generalizzazione di Università in MedioPiccole e Medie non è né totale né esclusiva
 - le generalizzazioni non esclusive vengono tipicamente “corrette” aggiungendo esplicitamente l’entità “intersecata”
 - negli esempi di cui sopra, si avrà la generalizzazione di Università in Medie, Piccole e Grandi che è ora totale ed esclusiva, e quella di Università in Medie e Piccole, che è non totale ed esclusiva
- Slide 67: generalizzazione totale, freccia “piena”
 - Slide 68: generalizzazione non totale, freccia “non piena”