Sistemi di Elaborazione dell'informazione II

Corso di Laurea Specialistica in Ingegneria Telematica
II anno – 4 CFU
Università Kore – Enna – A.A. 2008-2009

Alessandro Longheu
http://www.diit.unict.it/users/alongheu
alessandro.longheu@diit.unict.it

Introduzione

Natura dei dati

- I DBMS relazionali sono utilizzati in numerose applicazioni di grande rilevanza, (sistemi informatici di banche e aziende), ma la maggior parte dei dati oggi disponibili in formato digitale non è rappresentabile sotto forma di relazioni.
- La produzione di grosse quantità di dati non relazionali si è intensificata nel tempo a causa della diffusione di Internet; questi dati hanno in generale caratteristiche differenti da quelle dei dati tipicamente gestiti tramite il modello relazionale.

Natura dei dati

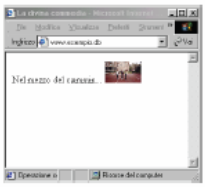
- I dati possono essere classificati in strutturati e non strutturati, ovviamente con una categorizzazione intermedia che prende il nome di semi-strutturati

Strutturati

Non Strutturati

id-pers	nome	cognome
0000001	Mario	Rossi
0000002	Giorgio	Verdi

id-pers	telefono
0000001	051 1234
0000001	333 3333



Natura dei dati

- I dati strutturati sono quelli caratterizzati da uno schema, quindi di fatto quelli gestiti dai DBMS classici
- Al versante opposto, i dati non strutturati sono completamente privi di schema e possono essere identificate due categorie:
  - Dati grezzi, ad esempio immagini
  - Dati senza schema, ad esempio porzioni di testo
- In una posizione intermedia, i dati semi-strutturati sono quelli per i quali esiste una struttura parziale, non sufficiente tuttavia per permetterne la memorizzazione e gestione da parte dei DBMS relazionali

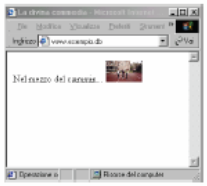
Natura dei dati

- Dati strutturati

id-pers	*nome*	*cognome*
0000001	Mario	Rossi
0000002	Giorgio	Verdi

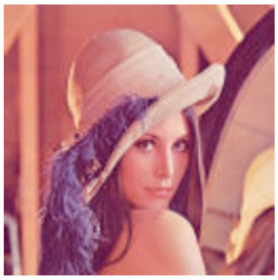
id-pers	*telefono*
0000001	051 1234
0000001	333 3333

DATI STRUTTURATI (SCHEMA)



Natura dei dati

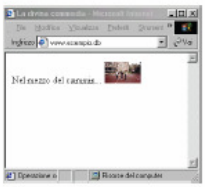
- Dati non strutturati grezzi



vu45s89gysJPGi
8gbyygsvs954gy
4598y9syg5vts9
4lygs98yg9s45y
g584gyt459gyg4
…

id-pers	*nome*	*cognome*
0000001	Mario	Rossi
0000002	Giorgio	Verdi

id-pers	*telefono*
0000001	051 1234
0000001	333 3333



RAW DATA

Natura dei dati

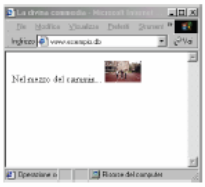
- Dati di testo non strutturato

E' molto meglio essere bello che buono; ma e' meglio essere buono piuttosto che brutto.
O.Wilde

L'ETERNO
Tra un fiore colto e l'altro donato
l'inesprimibile nulla
G.Ungaretti

id-pers	nome	cognome
0000001	Mario	Rossi
0000002	Giorgio	Verdi

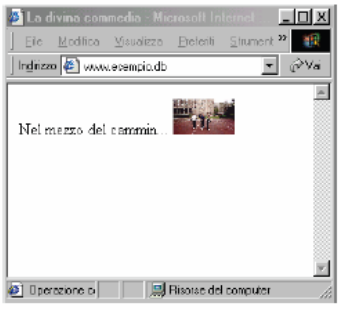
id-pers	telefono
0000001	051 1234
0000001	333 3333



DATI SENZA SCHEMA

Natura dei dati

- Dati semi-strutturati

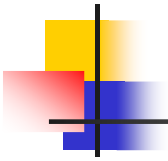



```
<html>
<title>La divina commedia</title>
Nel mezzo del cammin…
<img/></html>
```

id-pers	nome	cognome
0000001	Mario	Rossi
0000002	Giorgio	Verdi

id-pers	telefono
0000001	051 1234
0000001	333 3333

DATI CON STRUTTURA PARZIALE

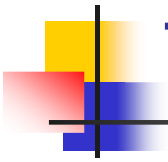


Natura dei dati

- Confronto:

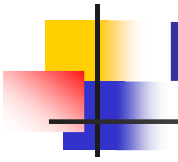
Relazionali	Semi-strutturati
Netta divisione tra schema e dati	Schema parziale e con caratteristiche simili ai dati
Basati sul concetto di insieme	Basati sul concetto di lista
Non ordinati	Ordinati
Non annidati	Annidati

Relazionali	Non-strutturati
Netta divisione tra schema e dati	Nessuno schema
Linguaggio di interrogazione	Linguaggio di ricerca
Modello booleano (correttez&completezza)	Modello basato su Ranking
Aggiornamenti ed interrogazioni parziali	Aggiornamenti totali



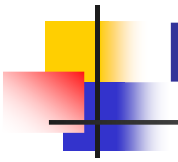
Trattamento dei dati

- Come e' noto, i dati strutturati sono trattati con le classiche tecniche previste dalla teoria dei DBMS
- I dati all'estrema destra di questa "classifica" sono detti dati non strutturati, e necessitano di un trattamento specifico. La disciplina che studia come manipolare questi dati è l'Information Retrieval.
- I dati nel mezzo sono detti dati semi-strutturati, e presentano caratteristiche sia dei dati strutturati che dei dati non strutturati.
- Uno dei linguaggi più utilizzati per la rappresentazione di dati semi-strutturati è XML. Il modello relazionale non è adatto a gestire questi tipi di dati.



Dati Semistrutturati

- Il formato principale per la rappresentazione di dati semistrutturati è XML.
- Esso può essere utilizzato sia per rappresentare dati strutturati, ad esempio allo scopo di scambiarli tra diverse applicazioni, che per rappresentare dati semi-strutturati, sfruttandone la flessibilità e la possibilità di indicare sia i dati che lo schema.
- Nel primo caso, i dati possono risiedere all'origine in un sistema relazionale, ed essere poi convertiti in XML; nel secondo caso, il modello relazionale non risulta essere particolarmente adatto alla gestione di questi dati.



Dati Semistrutturati

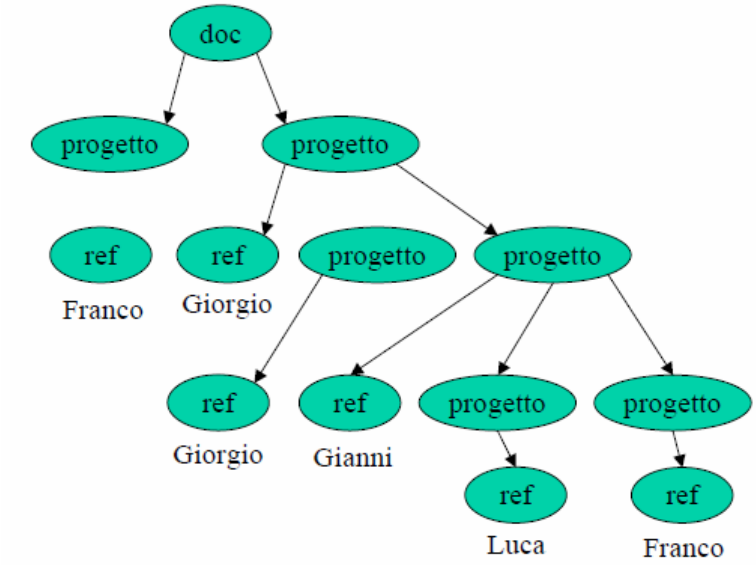
Esempio di file XML:

```
<doc>
  <progetto><ref>Franco</ref></progetto>
  <progetto>
        <ref>Giorgio</ref>
        <progetto><ref>Giorgio</ref></progetto>
        <progetto><ref>Gianni</ref>
              <progetto><ref>Luca</ref></progetto>
              <progetto><ref>Franco</ref></progetto>
        </progetto>
  </progetto>
</doc>
```

- E' possibile o conveniente rappresentare questi dati con DBMS relazionali?

Dati Semistrutturati

- Struttura ad albero del file XML:

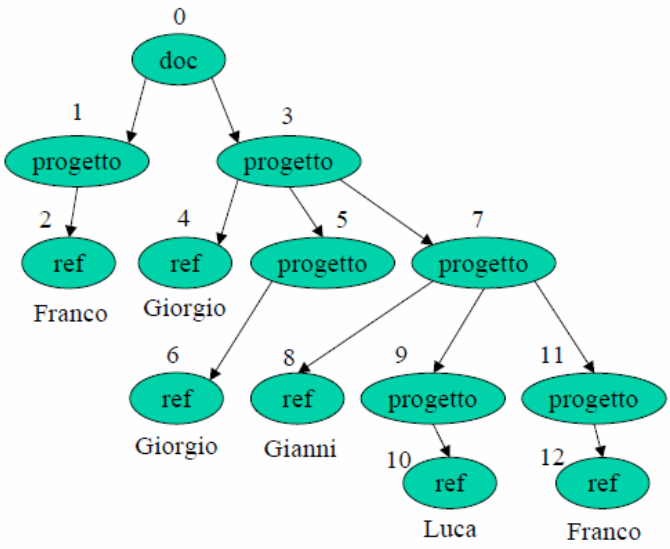


Dati Semistrutturati

- Volendo rappresentare l'albero in DBMS relazionale:

id	Nome
0	doc
1	progetto
2	ref
3	progetto
4	ref
5	progetto
6	ref
7	progetto
8	ref
9	progetto
10	ref
11	progetto
12	ref

id	figlio
0	1
0	3
1	2
3	4
3	5
3	7
5	6
7	8
7	9
7	11
9	10
11	12



Dati Semistrutturati

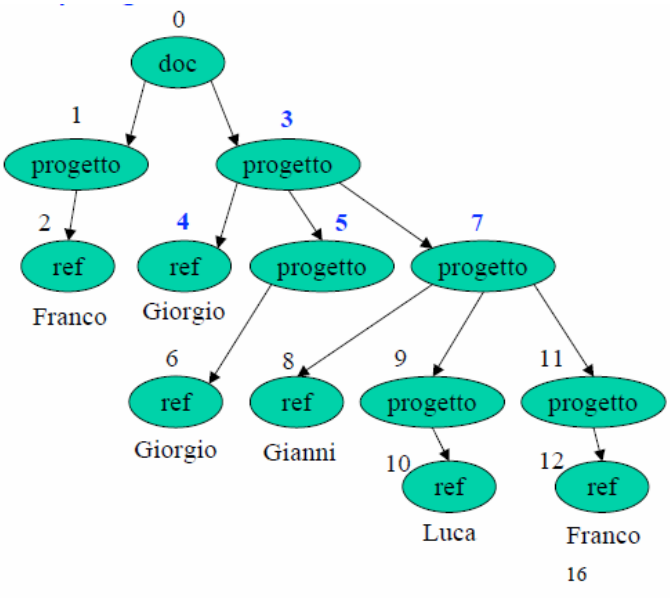
- La soluzione presentata ha alcuni limiti, dovuti al fatto che XML nasce per scambiare dati tra applicazioni e per rappresentare dati comprensibili anche da esseri umani. Le tabelle dell'esempio perdono questa caratteristica. Il modello dei dati e' quindi piu' complicato del formato originale.
- Alcune interrogazioni "ragionevoli" non si possono scrivere in SQL senza utilizzare la ricorsione, oppure possono risultare inefficienti, richiedendo più volte accesso alla stessa tabella. Ad esempio "Trova tutti i referenti che partecipano al secondo progetto".
- Occorre menzionare che questa strada, opportunamente migliorata, è stata percorsa dalla comunità scientifica ottenendo buoni risultati. Tuttavia, la tendenza attuale è quella di sviluppare sistemi specifici per XML.

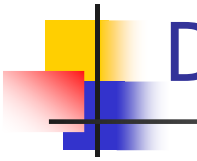
Dati Semistrutturati

- Query "Trova tutti i referenti che partecipano al secondo progetto"

id	Nome
0	doc
1	progetto
2	ref
3	progetto
4	ref
5	progetto
6	ref
7	progetto
8	ref
9	progetto
10	ref
11	progetto
12	ref

id	figlio
0	1
0	3
1	2
3	4
3	5
3	7
5	6
7	8
7	9
7	11
9	10
11	12



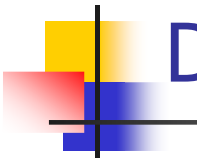


Dati Semistrutturati

- Esistono ancora altri limiti al modello relazionale:

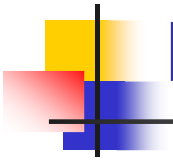
Autore	Titolo	Nascita	Testo
Dante Alighieri	La Divina Commedia	1265	Canto I Nel mezzo del…

- Il campo testo puo' anche contenere molti caratteri, senza alcuna struttura
- Per aggiungere altre informazioni occorre modificare la struttura della tabella.
- Se si vogliono scambiare i dati con altre applicazioni, oltre ai dati occorre inviare anche lo schema (ad esempio 1265 senza "nascita"), altrimenti essi potrebbero risultare incomprensibili.



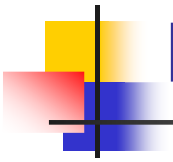
Dati Semistrutturati

1. La struttura è irregolare o parziale.
2. Lo schema è costruito a posteriori (data guide).
3. Lo schema è molto ampio.
4. Lo schema evolve rapidamente.
5. Le differenze tra schema e dati non sono significative.
6. Lo schema viene modificato.
7. Lo schema viene comunicato insieme ai dati.
8. Lo schema non impone vincoli inappellabili.
9. Le interrogazioni riguardano anche lo schema.
10. Nel caso di XML, sono rilevanti l'ordine e l'annidamento reciproci dei dati.
11. Essendo ordinati, i dati sono rappresentati tramite liste.



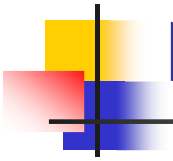
Dati Non strutturati

- In XML e nei dati semi-strutturati, lo schema ha caratteristiche particolari, ma è comunque presente. Ad esempio, una query in XPath su un documento XML accede tipicamente ai tag degli elementi, cioè allo schema.
- Se lo schema non è presente, come nel caso di oggetti multimediali e file di solo testo narrativo, le modalità di gestione di questi dati cambiano significativamente. La disciplina principale che studia questi dati si chiama Information Retrieval.
- Dati senza schema, o di cui tipicamente non si utilizza lo schema, sono di grandissima importanza: basti pensare a Internet e ai motori di ricerca, che sono per lo più sistemi di Web Information Retrieval.



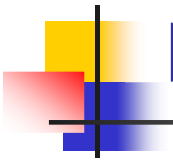
Dati Non strutturati

- Nonostante linguaggi di interrogazione più complessi siano stati proposti, nella maggioranza dei casi le interrogazioni su dati non strutturati (più che altro su dati testuali) sono molto semplici, solitamente composte da elenchi di parole chiave, ad esempio
- "Restituire i documenti che contengono la parola natura"
  e non…
- Select Nome, Count(distinct progetto), Sum(mesi) From Persona Natural Join Assegnazione Where Nome like 'M%' AND Eta > 40 Group by Nome



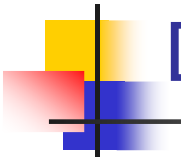
Dati Non strutturati

- Nelle interrogazioni a database relazionali, le interrogazioni esprimono requisiti precisi, e ogni tupla del risultato soddisfa pienamente quei requisiti. La costruzione della risposta a un'interrogazione relazionale segue dunque un modello booleano: una tupla è presente oppure non è presente nel risultato.
- Data la natura delle interrogazioni, data la grande quantità di possibili risposte, e dato che diversi documenti possono rispondere più o meno bene ai requisiti espressi nell'interrogazione, nell'information retrieval un modello booleano spesso non è utilizzabile.



Dati Non strutturati

- Dato il gap tra dati e informazioni, dovuto all'ambiguità dei dati, spesso non si riesce a determinare con precisione se un risultato è completamente o per nulla rilevante.
- I risultati vengono quindi ordinati per gradi di rilevanza (vedi ad esempio Google), e l'utente ammette possibili "errori".
- Poiché non è solitamente possibile ritornare risultati corretti e completi, come avviene invece nel caso dei dati strutturati, vi sono metriche per descrivere la bontà di un risultato.
- Ad esempio:
- "Restituire i documenti che hanno come argomento la natura."
- Una risposta valida (?): "L'ETERNO Tra un **fiore** colto e l'altro donato l'inesprimibile nulla (G.Ungaretti)"



Dati Non strutturati

- Dagli esempi mostrati precedentemente emerge come le caratteristiche dei dati, delle interrogazioni e dei risultati delle stesse siano significativamente diverse da quelle riscontrate nei sistemi relazionali.
- Oltre all'ordinamento per rilevanza, il risultato di una query su dati non strutturati tipicamente non prevede la manipolazione dei dati, ma solo la selezione di alcuni di essi.
- Anche in questo caso, come e più che per i dati semi-strutturati, vi è dunque necessità di utilizzare modelli e sistemi differenti.
- Si noti che nei principali sistemi per la gestione delle basi di dati relazionali sono state integrate già da tempo funzionalità derivate dall'Information Retrieval, come l'indicizzazione di colonne contenenti solo testo (CLOB) o colonne per dati multimediali (BLOB).