

Archivi e loro organizzazioni

1. Concetti introduttivi

1.1 Concetto di archivio

In informatica il concetto di archivio coincide in gran parte con quello già posseduto da ciascuno di noi: un archivio è un insieme di registrazioni (o *records*) ciascuna delle quali è costituita da un insieme prefissato di informazioni elementari dette **attributi** (o *campi*). Indicativamente si può supporre che tali informazioni descrivano in qualche modo le *proprietà* interessanti di un *oggetto* o **entità**.

Si può pensare ad un archivio come ad una **tabella** le cui righe rappresentano le registrazioni e le cui colonne rappresentano sequenze di uno stesso attributo.

Nella figura 1.1 è riportato un semplice esempio di archivio relativo ad un gruppo di studenti di una scuola:

| <i>Matricola</i> | <i>Cognome</i> | <i>Nome</i> | <i>DataDiNascita</i> | <i>LuogoDiNascita</i> | <i>Sesso</i> |
|------------------|----------------|-------------|----------------------|-----------------------|--------------|
| 100 | Rossi | Marco | 20-09-1990 | Roma | M |
| 110 | Verdi | Maria | 06-11-1991 | Latina | F |
| 175 | Gialli | Aldo | 25-07-1990 | Latina | M |
| 183 | Neri | Marco | 13-12-1989 | Roma | M |
| 196 | Bianchi | Rosa | 30-09-1990 | Latina | F |

Figura 1.1 Archivio STUDENTI

1.2 Intensione ed Estensione di un attributo

Parlando genericamente di attributo occorre sempre saper distinguere fra *nome* o **intensione** dell'attributo e *valore* o **estensione** dell'attributo. Ad esempio, con riferimento all'archivio STUDENTI l'intensione di un attributo è *Sesso*. La sua estensione può essere 'M' o 'F'.

1.3 Chiavi primarie e chiavi secondarie

Un sottoinsieme degli attributi di un record viene detto **chiave**. Anche in questo caso occorre sempre saper distinguere fra intensione della chiave (l'insieme dei nomi degli attributi) e sua estensione (l'insieme dei valori degli attributi).

Una **chiave primaria** è una chiave che *individua al più una registrazione* fra quelle contenute nell'archivio: il suo valore può essere posseduto da una sola registrazione, cioè *individua un'entità*. Nel caso dell'archivio STUDENTI, una chiave primaria potrebbe essere {*Matricola*} oppure {*Cognome, Nome, DataDiNascita*}. Facendo l'ipotesi che tutti i records siano diversi una chiave primaria esiste sempre: essa è al limite costituita da tutti gli attributi.

Una chiave non primaria è detta **chiave secondaria**: il suo valore *individua, in generale, più di una registrazione*. Ad esempio, con riferimento all'archivio STUDENTI, si può osservare che $\{Nome\}$, $\{LuogoDiNascita\}$, $\{Sesso\}$, $\{LuogoDiNascita, Sesso\}$ sono tre possibili chiavi secondarie.

Una chiave secondaria *selettiva* è una chiave cui è associato un numero relativamente basso di registrazioni come ad esempio $\{Nome\}$ nell'archivio STUDENTI.

Nonostante possano esistere più chiavi primarie *candidate*, è sempre bene scegliere come chiave *primaria di riferimento* quella costituita da un solo attributo. Spesso, nell'ambito dei sistemi informativi, si introduce artificialmente un attributo che svolge unicamente il ruolo di chiave *primaria di riferimento* come ad esempio l'attributo Matricola nell'archivio STUDENTI. Salvo diversa precisazione, parlando genericamente di chiave, sottintenderemo la chiave primaria di riferimento. Faremo inoltre uso della notazione $rec(K)$ per indicare la registrazione avente chiave K .

2. Operazioni e interrogazioni su archivi

Nell'ambito della programmazione, da un punto di vista astratto, si può cercare di vedere ogni archivio come una variabile di tipo *archivio*. Questo, come tutti i tipi di dato, è individuato da un insieme di precise operazioni definite sui dati di tipo archivio.

Non esiste un generale accordo su quali debbano essere esattamente tali operazioni, ma in linea di massima su ogni archivio dovrebbe essere possibile compiere almeno le seguenti operazioni:

- 1) **Creazione** di un archivio vuoto.
- 2) **Inserimento** di un record con chiave K .
- 3) **Reperimento** di un record di chiave K .
- 4) **Aggiornamento** (di campi non chiave) di un record di chiave K .
- 5) **Cancellazione** di un record di chiave K .
- 6) **Visita** a tutti i records dell'archivio.

Le operazioni 2, 4 e 5 vengono generalmente classificate come **operazioni di modifica** e richiedono di fatto una ricerca preventiva (operazione 3). Ad esempio, l'operazione di inserimento di uno studente nell'archivio STUDENTI di figura 1.1 deve essere preceduta da una ricerca: il nuovo studente può essere inserito solo se nell'archivio non esiste già uno studente con la stessa matricola.

Il **reperimento** di un record consente di sapere se esiste un certo record e, in caso affermativo, di conoscerne tutti gli attributi.

La modifica ad un campo chiave può essere condotta solo cancellando prima il vecchio record e inserendo poi quello nuovo.

L'operazione di visita è essenziale in tutti quei casi in cui occorre applicare una certa elaborazione a tutti o parte dei records dell'archivio (si pensi ad una stampa o ad un aggiornamento). A livello elementare, l'operazione di

visita viene effettuata in base a due operazioni: una di accesso al "primo" record dell'archivio e una di accesso al record "successivo" all'ultimo cui si è fatto accesso.

2.1 Ordini di visita

Si distinguono due tipi di ordini di visita:

- a) visita non ordinata (visita in ordine qualsiasi)
- b) visita ordinata

La visita ordinata viene effettuata per valori crescenti o decrescenti di un attributo per il quale sia definita una relazione d'ordine totale (es. attributi di tipo numerico o stringa). Più in generale la visita ordinata può essere definita per una chiave qualsiasi, cioè formata da più di un attributo. In questi casi la relazione di ordinamento è definita dall'ordinamento lessicografico sulla ennupla ordinata costituita dagli attributi che compongono la chiave.

Ad esempio, considerando un ordinamento sulla chiave {Cognome, Nome} nell'archivio STUDENTI di figura 1.1, la relazione d'ordine

$$(\mathit{cognome1}, \mathit{nome1}) \leq (\mathit{cognome2}, \mathit{nome2})$$

significa

$$\mathit{cognome1} < \mathit{cognome2} \text{ o } \mathit{cognome1} = \mathit{cognome2} \text{ e } \mathit{nome1} \leq \mathit{nome2}.$$

La visita ordinata comporta un'operazione di ordinamento a meno che l'archivio non sia stato opportunamente implementato in modo da poter essere visitato in modo ordinato. Ciò accade molto spesso nei confronti della chiave primaria.

In effetti, parlando genericamente di *archivi ordinati*, si intende che è possibile eseguire la visita ordinata secondo valori crescenti (o decrescenti) della chiave senza ricorrere ad una procedura di ordinamento.

2.2 Interrogazione dell'archivio

L'operazione di reperimento di un record con chiave K è solo un caso particolare di **interrogazione** o **query** dell'archivio. Una sua generalizzazione presa frequentemente in considerazione è la seguente:

ricercare tutti i records i cui attributi soddisfano un predicato P

Il **predicato P** esprime *la condizione* che deve essere soddisfatta da un sottoinsieme degli attributi dei records; può essere **semplice** o **composto**.

Un **predicato semplice** mette in relazione il nome di un attributo con un dato valore e viene espresso nella forma:

<nome attributo> <oprel> <valore attributo>

dove <oprel> è uno dei sei operatori relazionali: =, ≤, ≥, ≠, <, >.

Esempi: LuogoDiNascita='Roma', Sesso='M', ecc.

Un **predicato composto** si ottiene da quelli semplici o composti mediante applicazione degli operatori booleani **OR**, **AND**, **NOT**.

Esempi:

LuogoDiNascita='Roma' AND Sesso='M', NOT LuogoDiNascita='Roma' AND Sesso='F', ecc.

3. Fattori che influenzano la scelta dell'organizzazione di un archivio

Con il termine **organizzazione di un archivio** si intende sia il *tipo di rappresentazione (struttura dati)* su memoria di massa di un archivio, sia le *procedure di base* per l'accesso ad un record (metodo di accesso). Sono state analizzate e proposte numerose organizzazioni le più note delle quali saranno esaminate più avanti.

In generale ogni organizzazione tende a rendere più efficienti alcune operazioni a discapito di altre. Per questo motivo è necessario conoscere i fattori che possono influenzare la scelta di un'organizzazione piuttosto di un'altra. Ecco un breve elenco dei fattori più rilevanti:

- 1) Tipi di operazioni previste e loro frequenza d'uso
- 2) Tempi e metodi di elaborazione
- 3) Frequenza di riferimento
- 4) Struttura dei records
- 5) Dimensione ed espansibilità dell'archivio
- 6) Tipo e dimensione del supporto di memoria di massa
- 7) Integrità e ripristino dell'archivio
- 8) Costi ed ambienti di produzione
- 9) Vincoli imposti da riferimenti
- 10) Visibilità del progettista.

3.1 Tipi di operazioni previste e loro frequenza d'uso.

Durante l'uso di un archivio, non tutte le operazioni vengono invocate con la stessa frequenza o vengono affatto invocate. Ad esempio si distingue spesso fra **archivio dinamico** e **archivio statico** a seconda che l'archivio sia o meno soggetto a frequenti inserimenti e/o cancellazioni.

In alcune organizzazioni le operazioni di inserimento e di cancellazione possono degradare l'efficienza della ricerca al punto che si può rendere necessaria una **periodica riorganizzazione** dell'archivio.

Se esiste l'esigenza di reperire frequentemente i records in base ad un particolare attributo può essere conveniente tenerne conto nell'organizzazione dell'archivio gestendo delle *strutture di accesso per chiave secondaria* (files di appoggio all'archivio principale).

Se sono previste frequenti interrogazioni per campo di variabilità di una chiave (es. ricercare tutte le automobili di prezzo compreso fra 8.000,00 euro e 13.000,00 euro) può essere conveniente ricorrere ad una organizzazione ordinata per quella chiave.

3.2 Tempi e metodi di elaborazione.

Se tutte le operazioni avvengono in *modo interattivo* il tempo medio di risposta deve essere molto breve (dell'ordine di pochi secondi). A seconda delle applicazioni potranno essere imposti anche dei vincoli sulla variabilità del tempo di risposta intorno al suo valor medio e ciò potrà influire sui parametri o sulla scelta dell'organizzazione. Questa sarà generalmente diversa da quella che conviene adottare nei casi in cui non esistono interrogazioni interattive e, più in generale, quando esiste un'ampia tolleranza sul tempo di risposta. In quest'ultimo caso le richieste di modifica o di ricerca possono essere raccolte in un apposito file e applicate periodicamente tutte in una volta (elaborazione non interattiva o *off-line*). Ciò sarà ancor più conveniente in presenza di un elevato **rapporto di attività** definito come il rapporto fra numero di records da elaborare e numero di records presenti nell'archivio (con riferimento ad una certa unità di tempo).

Nel valutare la complessità in tempo di un'organizzazione si considera come operazione rilevante quella di accesso in memoria di massa. Salvo rare eccezioni si possono infatti ritenere trascurabili i tempi di elaborazione in memoria centrale.

3.3 Frequenza di riferimento

In genere non tutti i records vengono riferiti con la stessa frequenza. Molto spesso vige anzi la regola dell'"80-20", cioè l'80% degli accessi vengono riferiti al 20% dei records presenti nell'archivio. Si può tener conto di questo aspetto nel progetto dell'organizzazione cercando di rendere più facilmente accessibili i records che hanno maggiore probabilità di essere riferiti. Ciò può avvenire in modo dinamico; un criterio potrebbe essere quello di migliorare la posizione di un record nell'archivio a scapito di altri ogni n volte che viene riferito.

3.4 Struttura dei records

I records possono essere a *lunghezza fissa* o *variabile*. Quest'ultimo caso si presenta tipicamente quando esistono degli attributi di tipo *string* (*stringa illimitata*) oppure in presenza di *gruppi di attributi ripetuti* un numero arbitrario di volte. Ad esempio, in un archivio anagrafico, nel record che descrive una persona potrebbe esserci un gruppo ripetuto ciascuno dei quali descrive il nome, la data e il luogo di nascita di ciascuno dei suoi figli. Come si può intuire, i records a lunghezza variabile possono richiedere l'adozione di organizzazioni diverse da quelle usate per i records a lunghezza fissa. Tuttavia, anche se la lunghezza è fissa, il suo valore in bytes o il rapporto fra lunghezza della chiave e lunghezza del record può influire sul tipo di organizzazione da scegliere.

In molti casi, invece del valore dell'attributo può essere più conveniente mettere un riferimento (un puntatore) ad un'area di memoria di massa contenente tale valore. Ciò è utile per realizzare la cosiddetta *condivisione dei dati* (o *data sharing*).

3.5 Dimensione ed espansibilità dell'archivio

In generale il ricorso ad organizzazioni sofisticate è tanto più giustificato quanto più grande è l'archivio. Occorre comunque valutare il fatto che gli archivi possono crescere e che potrebbe essere sconveniente dover cambiare successivamente l'organizzazione e/o il supporto di memoria secondaria.

Nella scelta di un'organizzazione occorre anche trovare un giusto compromesso fra spazio e tempo: avremo infatti modo di vedere come valga spesso la regola secondo cui per guadagnare in velocità occorre pagare un certo costo in spazio. Tuttavia, ricorrendo alle *tecniche di compressione* (codifica corta) dei dati si può spesso ottenere il duplice risultato di guadagnare in velocità riducendo lo spazio occupato.

3.6 Tipo e dimensione del supporto di memoria di massa

Se il supporto di memoria disponibile è ad *accesso sequenziale* (nastri) vengono necessariamente escluse tutte le organizzazioni basate sull'*accesso diretto*. Nel caso dei dischi può essere utile considerare il fatto che le testine siano fisse o mobili. Se la dimensione del supporto è critica può non essere possibile adottare alcune organizzazioni veloci che richiedono dello spazio aggiuntivo rispetto a quello strettamente necessario per memorizzare i records dell'archivio.

3.7 Integrità e ripristino dell'archivio

Occorre valutare le conseguenze ed i costi dovuti alla distruzione di una parte delle informazioni dell'archivio in seguito a cause accidentali o per errore di programmazione. A tale proposito si osservi che se l'archivio è rappresentato mediante uso di puntatori, l'alterazione di un puntatore può rendere inaccessibile una parte consistente di informazioni. Nell'organizzare l'archivio occorrerà comunque prevedere i necessari accorgimenti per permettere la *ricostruzione dell'archivio a partire dall'ultima copia salvata* su nastro o su qualsiasi supporto rimovibile. La tecnica generalmente adottata consiste nel memorizzare su un apposito file (sequenziale) la *storia delle modifiche* avvenute dopo l'ultimo salvataggio. La ricostruzione può avvenire applicando successivamente tali modifiche all'ultima copia salvata.

3.8 Costi ed ambiente di produzione

In generale il progetto e la messa a punto di un'organizzazione per archivi comporta un certo numero di iterazioni progettuali e sperimentali che si traducono in un costo non trascurabile. Ciò spiega il fatto che in molti casi la scelta dell'organizzazione sia pesantemente condizionata dall'ambiente di produzione del software applicativo. In effetti accade spesso che si sia disposti ad accettare delle prestazioni mediocri sia in tempo che in spazio

piuttosto che rinunciare a delle organizzazioni già collaudate e disponibili da tempo. A tale riguardo molta influenza possono avere i linguaggi ed il Sistema Operativo adottati dall'ambiente. Occorre comunque tenere presente che una scelta oculata dell'organizzazione può essere ampiamente ripagata da dei costi di gestione dell'archivio molto contenuti rispetto a quelli necessari per una scelta più immediata ma anche meno felice nei confronti dell'efficienza in tempo e spazio.

3.9 Vincoli imposti da riferimenti

Può accadere che i records dell'archivio siano "puntati" (pinned records), cioè che il loro indirizzo sia utilizzato da qualche struttura dati presente nello stesso ambiente. Ciò impone di fatto la necessità che il record non possa essere rimosso dalla sua posizione almeno fino a quando non c'è la certezza che non sia più puntato. Questo vincolo introduce qualche problema in più rispetto al caso in cui i records non sono puntati (unpinned records) ed influenza il tipo di organizzazione che è possibile adottare. Una soluzione a questo problema consiste nel sostituire i puntatori fisici con dei puntatori simbolici costituiti dalle chiavi primarie dei records. Ciò richiede degli accessi supplementari ma si è liberi di scegliere l'organizzazione che si ritiene più opportuna.

3.10 Visibilità del progettista.

L'organizzazione non può essere scelta prescindendo dal linguaggio utilizzato per definirla. Il linguaggio stabilisce in pratica il livello di visibilità in cui si pone il progettista dell'archivio. Le prestazioni di un'organizzazione possono infatti essere profondamente diverse a seconda che il progettista abbia la visibilità di un linguaggio ad alto livello, del SO o dell'implementazione delle aree di memoria secondaria.

QUESTIONARIO DI AUTOVALUTAZIONE

1. Che cosa si intende per "organizzazione di un archivio"?

[a] Si intende sia la struttura dati in memoria di massa, sia le procedure di base per l'accesso ad un record.

[b] Si intende la sola struttura dati in memoria di massa atta a rappresentare le informazioni.

[c] Si intende l'insieme delle procedure di base che consentono di leggere e scrivere dati in memoria di massa.

[d] Si intende la sola struttura dati in memoria RAM atta a rappresentare le informazioni di interesse.

[e] Si intende sia la struttura dati in memoria RAM atta a rappresentare le informazioni, sia la struttura dati in memoria di massa.

2. Quale dei seguenti fattori è il meno rilevante nella scelta dell'organizzazione di un archivio?

[a] Tipo di operazioni previste e loro frequenza d'uso.

[b] Tipo e dimensione del supporto di memorizzazione.

[c] Tipo e velocità del microprocessore.

[d] Tempi e metodi di elaborazione.

[e] Costi ed ambiente di produzione.

3. Quale delle seguenti operazioni influenza maggiormente le prestazioni di un archivio?

[a] L'operazione di cancellazione.

[b] L'operazione di inserimento.

[c] L'operazione di modifica.

[d] L'operazione di ricerca.

[e] L'operazione di visita non ordinata di tutti i records dell'archivio.

4. Che cosa si intende per "intensione" di un attributo di un record?

[a] Il solo valore corrente dell'attributo.

[b] Il solo identificatore dell'attributo.

[c] Il valore corrente dell'attributo ed il suo identificatore.

[d] L'identificatore dell'attributo ed il tipo di valori che esso può assumere.

[e] Il numero massimo di valori che l'attributo può assumere.

5. Che cosa si intende per "estensione" di un attributo di un record?

[a] Il solo valore corrente dell'attributo.

[b] Il solo identificatore dell'attributo.

[c] Il valore corrente dell'attributo ed il suo identificatore.

[d] L'identificatore dell'attributo ed il tipo di valori che esso può assumere.

[e] Il numero massimo di valori che l'attributo può assumere.

6. Che cos'è una "chiave primaria"?

[a] Un sottoinsieme qualsiasi degli attributi di un record.

[b] Un sottoinsieme qualsiasi degli attributi di un record che consente di individuare almeno una registrazione all'interno dell'archivio.

[c] Un sottoinsieme qualsiasi degli attributi di un record che consente di individuare solo una registrazione all'interno dell'archivio.

[d] Un sottoinsieme qualsiasi degli attributi di un record che consente di individuare al più una registrazione all'interno dell'archivio.

[e] Un sottoinsieme qualsiasi degli attributi di un record che consente di individuare più di una registrazione all'interno dell'archivio.

7. Che cos'è una "chiave secondaria"?

[a] Un sottoinsieme qualsiasi costituito da almeno due attributi di un record.

[b] Un sottoinsieme qualsiasi degli attributi di un record che consente di individuare almeno una registrazione all'interno dell'archivio.

[c] Un sottoinsieme qualsiasi degli attributi di un record che consente di individuare solo una registrazione all'interno dell'archivio.

[d] Un sottoinsieme qualsiasi degli attributi di un record che consente di individuare al più una registrazione all'interno dell'archivio.

[e] Un sottoinsieme qualsiasi degli attributi di un record che consente di individuare più di una registrazione all'interno dell'archivio.