File ed Archivi

Prima parte

Gestione dei File in ANSI C

- Nozione di file byte stream
- Dichiarazioni e definizioni
- Operazioni di base e procedure di gestione

Gestione dei File in ANSI C++

- Nozione di file byte stream
- Dichiarazioni e definizioni
- Operazioni di base e procedure di gestione

Archivi

- Nozioni di archivio ed attributo
- Chiave candidata, primaria e secondaria
- Operazioni di base e procedure di gestione
- Fattori di scelta di un'organizzazione

Organizzazioni sequenziali

- Caratteristiche
- Accesso sequenziale
- Accesso diretto
- Ricerca
- Inserimento
- Cancellazione
- Aggiornamento

Organizzazioni Hash

- Indirizzamento hash
- Requisiti di una funzione hash
- Risoluzione delle collisioni
- Indirizzamento aperto
- Concatenazione
- Valutazioni

Organizzazioni sequenziali con indice

- File indice e file primario
- Inserimenti
- Cancellazioni
- Indici multilivello
- Valutazioni

Organizzazioni a B-albero

- Alberi binari, AVL ed ABR
- Nozione di B-albero
- Ricerca di una chiave
- Visita completa
- Inserimento
- Cancellazione
- Valutazioni

Organizzazioni sequenziali con indice a B-albero

- Caratteristiche
- Valutazioni

Organizzazioni per chiavi secondarie

- Caratteristiche
- Organizzazione a liste multiple
- Organizzazione a liste invertite

Introduzione

Archivi in memoria di massa

- Costituiscono i modelli secondo cui sono organizzati i grandi "contenitori" di informazione
  - *Archivio anagrafico nazionale*
  - *Archivio cittadini italiani assistiti dal SSN*
  - *Archivio veicoli della motorizzazione civile*
  - *Archivio studenti iscritti all'ITIS "Kennedy"*
  - ...

Introduzione

Archivi in memoria di massa

- Le tecniche di implementazione sono le stesse impiegate nella realizzazione fisica delle basi di dati

Terminologia

- Archivio
- Registrazione (Record)
- Attributo (Campo)
- Entità
- Proprietà

Terminologia

- Entità → Registrazione
- Proprietà → Attributo

Rappresentazione sintetica di un archivio

(*Nominativo, Data di Nascita, Professione, Residenza*)

Rappresentazione tabellare di un archivio

Nominativo	Data di Nascita	Professione	Residenza
Bianchi Luca	*12/02/1963*	*Studente*	*Pordenone*
Rossi Marco	*26/11/1970*	*Ingegnere*	*Venezia*
Verdi Paolo	*5/05/1967*	*Avvocato*	*Trieste*

Terminologia

- Attributo
- Nome (intensione)
- Valore (estensione)

Chiave = sottoinsieme degli attributi di un record

Terminologia

- **Chiave Primaria** = *individua al più un record fra quelli contenuti nell'archivio. Ciò è come dire che ogni suo valore individua al massimo una sola entità.*

Se l'archivio è costituito da registrazioni tutte diverse, è allora vero che una chiave primaria esiste sempre. Al limite essa è costituita dall'insieme di tutti gli attributi.

Terminologia

- **Chiave Secondaria** = *chiave non primaria, ovvero individua, in generale, più di un record fra quelli contenuti nell'archivio.*

Chiave Secondaria

- **Selettiva** = ad ogni suo valore corrisponde un numero relativamente limitato di registrazioni presenti nell'archivio.
- **Non Selettiva** = ad ogni suo valore può corrispondere un numero anche molto elevato di registrazioni, al limite anche tutte quelle presenti.

Terminologia

- In un archivio possono esistere diverse chiavi primarie (disgiunte o meno). Si può ipotizzare però che esista un attributo particolare con funzione di chiave primaria di riferimento. Spesso, se tale attributo non esiste, lo si può facilmente creare.

- Qualche esempio
- Il **codice fiscale** è chiave primaria di riferimento per l'archivio anagrafico nazionale. **RSSLLN65B27G888S**
- Il **codice di riferimento assistito (C.R.A.)** è chiave primaria di riferimento per l'archivio degli iscritti ad usufruire del SSN. **3M7Q35**

Terminologia

- Qualche esempio
- Il **numero di targa** è chiave primaria di riferimento per l'archivio veicoli immatricolati presso la motorizzazione civile. DJ941PG
- Il **codice volo** è chiave primaria di riferimento per l'archivio internazionale dei voli registrati dalle compagnie aeree iscritte. AZ504
- Il **codice socio** è chiave primaria di riferimento per l'archivio soci iscritti al Club dei mangiatori di foglie di carciofo di Magonza (Germania). 087524

Operazioni su archivio

Ipotesi di lavoro:

- Nel seguito con il termine **chiave** intenderemo, tranne che in casi particolari, la chiave primaria di riferimento.
- Da un punto di vista astratto si può intendere un archivio come **una variabile di un tipo archivio**. Per questo, come per ogni altro tipo di dato, sono definite precise operazioni possibili sulle sue variabili, cioè su un determinato esemplare di archivio esistente.

Operazioni su archivio

Categorie di Operazioni

- Inizializzazione (I)
- Modifica (M)
- Interrogazione (Q)

Operazioni su archivio

Elenco delle Operazioni

- Creazione dell'archivio vuoto (I)
- Inserimento del record con chiave k (M)
- Ricerca del record con chiave k (Q)
- Aggiornamento (di campi diversi dalla chiave) del record con chiave k (M)
- Cancellazione del record con chiave k (M)
- Visita dell'intero archivio (Q)

Operazioni su archivio

- Le operazioni classificate “di modifica” richiedono in realtà una ricerca preventiva.
- La ricerca di un record consiste nel determinare se un certo record esiste o meno e, in caso affermativo, nel conoscerne tutti gli attributi
- La modifica di un campo chiave può essere realizzata mediante la seguente sequenza:
  - **Cancellazione del vecchio record (vecchia chiave)**
  - **Inserimento del nuovo record (nuova chiave)**
- **L’operazione di visita** dell’intero archivio è fondamentale in tutti i casi in cui occorre applicare una certa elaborazione a tutti i record di cui l’archivio è composto (o ad una parte di essi). Esempi di operazioni siffatte sono la stampa completa o l’aggiornamento.

Operazioni su archivio

- A livello elementare l'operazione di visita viene effettuata ricorrendo a due operazioni atomiche:
  - Accesso al primo record dell'archivio
  - Accesso al record successivo
- Esistono due tipi di visita:
  - Visita in ordine qualsiasi
  - Visita ordinata (per valori crescenti o decrescenti di un attributo)

Naturalmente il tipo dell'attributo deve essere caratterizzato da una relazione d'ordine totale (ad esempio un tipo integrale, carattere, stringa, ...)

- Ordinamento naturale
- Ordinamento secondo il codice di rappresentazione dei caratteri
- Ordinamento lessicografico
- ...

Operazioni su archivio

- Visita ordinata → necessità di una scansione ordinata
  - Operazione preventiva di ordinamento
  - Archivio già ordinato (per implementazione)

Questa seconda situazione è abbastanza frequente (con riferimento alla chiave primaria) per molte organizzazioni reali

- Archivio ordinato = è possibile effettuare una visita ordinata secondo valori crescenti o decrescenti della chiave senza ricorrere ad una procedura di ordinamento

Interrogazioni su archivio

- Ricerca di un record avente chiave k
  - si tratta di un caso particolare di interrogazione (query) su archivio.

- Ricerca di tutti i record aventi chiave k compresa fra k1 e k2
  - è un caso più generale e molto frequente
  - È denominata Range Query (interrogazione per campo di variabilità).
  - Se l'archivio non è ordinato è richiesta la visita completa

Esempio: "determinare tutti gli allievi dell'I.T.I.S. Kennedy con età compresa fra 16 e 18 anni"

Interrogazioni su archivio

- Ricerca di tutti i record i cui attributi soddisfano un predicato P
- Il predicato P può essere:
  - Semplice
  - Composto

Forma generale di un predicato semplice:

Attributo - operatore_relazionale - valore

Operatori relazionali (con riferimento al linguaggio C/C++):

== != < > <= >=

Interrogazioni su archivio

- Predicato composto: si ottiene mediante composizione di predicati semplici con gli operatori logici (booleani)

- Operatori logici (con riferimento al linguaggio C/C++):

  ! || &&

- Esempi:
  - **(Residenza == "Pordenone") && (Altezza >= 178 cm)**
  - **(Residenza == "Pordenone") || (Residenza != "Pordenone" && !((Età <16) || (Età >=18)))**

Interrogazioni su archivio

- Forma canonica di un predicato P: si ottiene mediante una disgiunzione di termini in cui ciascuno di questi ultimi risulta da una congiunzione di fattori; ciascun fattore non è altro che un predicato semplice.
- Ovvero:
- Fattore = predicato semplice
- Termine = F1 && F2 && F3 && … && Fn
- Predicato = T1 || T2 || T3 || … || Tm

Esempio di predicato composto in forma canonica:

- (Residenza == “Pordenone”) || (Residenza != “Pordenone” && Età >=16 && Età <18)

Fattori di scelta di un'organizzazione

- Organizzazione: tipo di implementazione di un archivio. Si intende con ciò indicare:
  - Tipo di rappresentazione in memoria di massa (strutture dati)
  - Procedure di base per l'accesso ad un record
- Non è indifferente la scelta di un'organizzazione piuttosto che un'altra!
- Ogni organizzazione proposta ed implementata comporta vantaggi e svantaggi.
- Si deve distinguere la complessità spaziale (l'occupazione di spazio nella MM) dalla complessità temporale (efficienza ovvero velocità di esecuzione delle principali operazioni richieste)

Fattori di scelta di un'organizzazione

È importante ricordare il problema della complessità degli algoritmi (vedi programma classe IV)

Complessità computazionale in senso temporale per operatività su archivi: si considera come operazione critica quella di accesso alla memoria di massa (accesso a file)

Fattori di scelta di un'organizzazione

- **Dinamicità (Volatilità):** si definisce dinamico o volatile un archivio soggetto a frequenti inserimenti e/o cancellazioni di record. Nel caso opposto si parla di Staticità (non volatilità).
- **Riorganizzazione**: si tratta dell'operazione che si rende necessaria quando la struttura di un archivio è stata pesantemente degradata in seguito a cancellazioni e inserimenti ripetuti (si tratta quindi di una necessità normale per gli archivi dinamici)
- **Rapporto di attività**: è il rapporto fra il numero di record da elaborare rispetto al numero totale di record presenti in archivio, con riferimento ad una determinata unità di tempo (ad es. ore, giorni, mesi, ecc)

Fattori di scelta di un'organizzazione

1. Tipi di operazioni previste e loro frequenza di utilizzo
   - Quali sono le operazioni che si effettueranno sull'archivio?
   - Quali di esse saranno effettuate con elevata frequenza e quali invece soltanto raramente?
   - Si tratterà di un archivio volatile o meno?
   - Saranno richieste pesanti riorganizzazioni periodiche?
   - Potranno essere interessanti ricerche anche per particolari chiavi secondarie? Saranno quindi auspicabili strutture di accesso per chiave secondaria?
   - Sono previste range query rispetto ad una certa chiave? È conveniente quindi mantenere l'ordinamento rispetto a quella chiave?

Fattori di scelta di un'organizzazione

1. Tempi e metodi di elaborazione
   - Le operazioni avranno luogo in modo interattivo? In questo caso il tempo di risposta dovrà essere brevissimo (dell'ordine del secondo). E la varianza del tempo di risposta?
   - Le operazioni verranno eseguite invece in modalità off-line? In questo caso il tempo di risposta non è significativo. Le richieste di modifica o di ricerca possono essere raccolte in un apposito file ed applicate periodicamente tutte in una volta. Questo sarà tanto più vantaggioso in presenza di un elevato rapporto di attività.

Fattori di scelta di un'organizzazione

1. Frequenza dei riferimenti
   - Non tutti i record vengono riferiti con identica frequenza
   - Molto spesso vale la regola dell'80/20: l'80% dei riferimenti è effettuato su un nucleo composto dal 20% dei record presenti in archivio
   - È il caso di privilegiare in qualche modo l'accesso a questo nucleo di record rispetto agli altri?

Si possono adottare anche metodi dinamici: un record "salta" più avanti di p (p>=1) posizioni ogni n (n>=1) riferimenti ad esso.

Fattori di scelta di un'organizzazione

1. Struttura dei record
   - Record a lunghezza fissa
   - Record a lunghezza variabile

Attributi del tipo stringa illimitata

Gruppi di attributi ripetuti N volte (N arbitrario)

Esempi:

Campo: figli (per un archivio anagrafico)

Campo: autoparco (per un archivio aziendale)

…

Record con varianti possono essere una soluzione pesante: è spesso preferibile la sostituzione con collegamenti (link) ad un altro archivio, strategia utile anche per ottenere la condivisione dei dati (data sharing)

Fattori di scelta di un'organizzazione

1. Dimensioni ed espandibilità dell'archivio
   - Il ricorso ad organizzazioni raffinate è tanto più giustificato quanto maggiori sono le dimensioni dell'archivio
   - Valutazioni sulle ipotesi di crescita dimensionale dell'archivio
   - Ricerca di un valido compromesso fra spazio e tempo, ovvero fra efficienza spaziale e temporale. In informatica vale sempre la regola: "ciò che si guadagna in spazio lo si paga in tempo e viceversa"
   - Possibile ricorso a tecniche di compressione e codifica breve

Fattori di scelta di un'organizzazione

1. Tipo e dimensione dei supporti di memoria di massa
   - Supporti ad accesso soltanto sequenziale (a nastro)
   - Supporti ad accesso anche diretto (dischi, ecc.)
   - Tecnologie a nastro molto efficienti attualmente: supporti estremamente capaci ma utilizzabili soltanto a scopo di backup
   - Tecnologie optoelettroniche avanzate: elevata capacità e prezzi contenuti. L'elaborazione utilizza però ancora pesantemente i supporti del tipo disco rigido (hard-disk)

Fattori di scelta di un'organizzazione

1. Sicurezza (Integrità e Ripristino)
   - Valutazione livello di rischio e conseguenze (costi!) per l'azienda
   - Eventi naturali ed eventi accidentali (guasti e rotture)
   - Eventi deliberati (intenzionali)

Necessità di tecniche che permettano la ricostruzione dell'archivio a partire dall'ultima copia salvata.

- Tecnologie RAID (non RAID 0!)
- Archivio storico delle modifiche: vengono memorizzate le variazioni intercorse dall'ultimo backup

Fattori di scelta di un'organizzazione

1. Costi ed ambiente di produzione
   - Molto spesso un'azienda è disposta ad accettare prestazioni mediocri sia in termini di efficienza che di occupazione spaziale piuttosto che rinunciare a organizzazioni in uso e collaudate in modo sicuro da lungo tempo
   - Ruolo ed influenza del SO e dei linguaggi e programmi applicativi di interfacciamento in uso
   - Una scelta attenta ed oculata dell'organizzazione verrà ripagata ampiamente da costi di gestione contenuti

Fattori di scelta di un'organizzazione

1. Vincoli imposti dai riferimenti
   - Può succedere che i record dell'archivio siano "puntati" (pinned) cioè che il loro indirizzo sia utilizzato da qualche struttura dati presente (ad esempio nei programmi applicativi di interfaccia). Questo esclude per esempio la possibilità di eliminazione (cancellazione fisica) del record. Questo costituisce un vincolo molto pesante.
   - Una soluzione può essere quella di sostituire i link "fisici" con link "simbolici" costituit dalle chiavi primarie di riferimento.

Fattori di scelta di un'organizzazione

1. Visibilità del progettista
   - Linguaggio disponibile per la realizzazione dell'archivio
   - Sistema operativo disponibile nell'ambiente di produzione

Modelli di memoria secondaria

- Gestione delle aree elementari su supporto di MM (disco)
- File Byte Stream (Byte file)
- File Paged (File impaginato, ovvero file visto come sequenza di blocchi o pagine di uguale dimensione fisicamente o logicamente contigui). In genere una pagina coincide con un settore del disco.

Archivi e livelli di astrazione

Organizzazioni Sequenziali

- Procedura di accesso al primo record
- Per ogni record (escluso uno) è definita un'operazione di accesso ad un altro record indicato come successivo. Questa operazione avviene in tempo costante, cioè indipendente dalla lunghezza del file o da altri fattori.
- Non esistono strutture ausiliarie (indici, algoritmi di ricerca e posizionamento, …)
- Metodo di accesso
  - Sequenziale
  - Diretto

Organizzazioni Sequenziali

- Metodo di accesso Sequenziale
  - Per accedere al record k-mo dopo aver fatto accesso al record i-mo è necessario un tempo T_{ik} dipendente da i e k
- Metodo di accesso Diretto
  - Il tempo per accedere a qualunque record è costante e indipendente dall'ultimo accesso effettuato. Questo metodo richiede l'utilizzo di un supporto di memoria adatto. L'accesso avviene generalmente mediante specificazione della posizione (primo record in posizione 0)

Organizzazioni Sequenziali

- Archivio ordinato

I record si presentano ordinati secondo i valori della chiave k

L'ordinamento facilita le operazioni di ricerca ma complica quella di inserimento

Organizzazioni Sequenziali Ricerca

- Accesso sequenziale (Archivio NON ordinato)
  - Caso di successo
    - Caso ottimo → 1 accesso
    - Caso pessimo → N accessi
    - Caso medio → (N+1)/2 accessi
  - Caso di insuccesso
    - Ricerca completa → N accessi

Nel caso di richieste NON equiprobabili è possibile migliorare l'efficienza memorizzando i record secondo l'ordine di probabilità di richiesta

Organizzazioni Sequenziali Ricerca

- Accesso sequenziale (Archivio ordinato)
  - Caso di successo
    - Caso ottimo → 1 accesso
    - Caso pessimo → N accessi
    - Caso medio → (N+1)/2 accessi
  - Caso di insuccesso
    - Caso ottimo → 1 accesso
    - Caso pessimo → N accessi
    - Caso medio → (N+1)/2 accessi

Organizzazioni Sequenziali Ricerca

- Accesso sequenziale
- (Archivi ordinati e NON ordinati)
- Nel caso di file impaginati i costi si riducono di un fattore NR pari al numero medio di record contenuti in una pagina

Organizzazioni Sequenziali Ricerca

- Accesso diretto (Archivio NON ordinato)
  - Caso di successo
    - Caso ottimo → 1 accesso
    - Caso pessimo → N accessi
    - Caso medio → (N+1)/2 accessi
  - Caso di insuccesso
    - Ricerca completa → N accessi

Nel caso di richieste NON equiprobabili è possibile migliorare l'efficienza memorizzando i record secondo l'ordine di probabilità di richiesta

Organizzazioni Sequenziali Ricerca

- Accesso diretto (Archivio ordinato)
- Si possono applicare algoritmi altamente efficienti (ricordare le nozioni apprese nella classe IV relativamente allo studio della complessità computazionale degli algoritmi!)
- Ricerca Binaria (o Dicotomica)
- Ricerca Interpolata

Organizzazioni Sequenziali Ricerca

- Accesso diretto (Archivio ordinato)
  - Caso di successo (Ricerca Binaria)
    - Caso ottimo → 1 accesso
    - Caso pessimo → $\log_2 N$ accessi
  - Caso di insuccesso
    - $\log_2 N$ accessi

Organizzazioni Sequenziali Inserimento

- Archivi NON ordinati
- L'inserimento di un nuovo record può avvenire
- In fondo al file
- Al posto di un record cancellato logicamente
- In un file secondario appositamente previsto

Archivi ordinati con accesso sequenziale

Procedura tipica:

- Raccolta dei record da inserire in un file separato
- Rigenerazione periodica dell'archivio mediante Merge con il file degli inserimenti previa operazione di Sort di quest'ultimo

Organizzazioni Sequenziali Inserimento

- Archivi ordinati con accesso diretto
- Medesima procedura vista nel caso dell'accesso sequenziale

- Archivi ordinati impaginati ad accesso diretto
- Tecnica della gestione in overflow:
- Free spaces (spazio lasciato libero all'atto della generazione iniziale dell'archivio, disponibile per i nuovi inserimenti)
- Area di overflow (file separato opportunamente previsto per ospitare gli overflow di pagine sature)
- Riorganizzazione quando:
  - Area di overflow esaurita
  - Prestazioni degradate nella ricerca

Organizzazioni Sequenziali
Inserimento

- Soluzioni possibili per l'organizzazione dell'area di overflow:
- Area unica (metodo delle liste di overflow)
  - i record di overflow di ciascuna pagina sono collegati ordinatamente a lista
  - Il link di accesso alla lista è contenuto nella pagina stessa
  - Tutte le liste "convivono" nella stessa area associata all'archivio principale (primario)
- Area distribuita
  - Tecnica dello splitting di pagina: se una pagina primaria è satura, ne viene allocata una di overflow nell'area secondaria e vi vengono trasferiti metà dei record presenti nella pagina satura, rispettando l'ordinamento. Ne nascono liste di overflow fra pagine. I link di accesso a tali liste sono contenuti nelle pagine primarie

Organizzazioni Sequenziali
Cancellazioni e Aggiornamenti

- Accesso sequenziale
  - Riscrittura completa dell'intero file modificato

In pratica si raccolgono tutte le modifiche in file detti differenziali. La riorganizzazione dell'intero archivio si effettua:

- Periodicamente
- Lunghezza eccessiva file differenziale e quindi elevato degrado dell'efficienza nelle ricerche

Organizzazioni Sequenziali
Cancellazioni e Aggiornamenti

- Accesso diretto (aggiornamento)
  - Riscrittura del solo record convolto

Accesso diretto (cancellazione)

- File non impaginato
  - Cancellazione logica mediante “flag di cancellazione”
  - Cancellazione fisica al momento della riorganizzazione
- File impaginato
  - Cancellazione fisica possibile anche se un poco più pesante nel caso in cui coinvolga anche l’area di overflow. Se il rapporto di occupazione rischia di abbassarsi troppo è utile la procedura di riorganizzazione

FINE