Appunti Database

Brendon Mendicino

October 6, 2022

Contents

1 Introduzione

KDD: Knowledge Discovery from Data

Tecniche di data mining.

Regole di associazione: usate per trovare delle relazioni frequenti all'interno del database. Ad esempio: chi compra pannolini compra anche birra, il 2% delle transazioni contegono entrambe gli oggetti, il 30% delle transaioni che contengono pannalini contengono anche birra.

Grazie alle regole di associazione si possono fare dei tipi di analisi come la basket analisys, ma puo essere utile anche per le raccomanadazioni.

Classificazione: i classificatori predicono etichette discrete, esempio: nella posta elettronica alcune mail vengono demoninati spam. La classificazione è definisce un modello per definire le predizioni, a volte non è sempre possibile creare dei modelli interpretabili ovvero dare una ragione per una determinata scelta.

Clustering: gli algoritmi danno gruppi di oggetti, senza però dare delle motivazioni.

...

- Fatto: fenomeno di studio;
- Misure: attributi del fatto;
- Dimensioni: tabelle collegate al fatto;

Schema a stella:

Snoflawke scheme: si esplicitano le dipendenze funzionali, questo però comporta un aumento delle operazioni di join.

Nella realtà lo snowflake è raramente utilizzato, il motivo è che il costo delle join può diventare oneroso. Un caso di utilizzo dello snowflake è quando si hanno dei dati condivisi.

Archi multipli:

Dimensioni degeneri: sono delle dimensioni con un solo attributo, questo si perchè nello stato attuale non si hanno delle specifiche per quell'attributo ma nel futuro si potrebbe facilmente estendere. Un'altra soluzione potrebbe essere un push down delle dimensioni degeri nella tabella dei fatti.

Junk Dimension: si può creare una dimensione che contenga tutte le dimensioni degeneri, le informazioni sono collegate semanticamente, è anche possibile unire delle informazioni scorrelate ma non è una scelta poco corretta, una soluzione potrebbe essere avere più junk dimensions.

2 Analisi

Sfruttando solo l'SQL è molto difficile fare delle analisi su un dw, infatti volendo calcolare delle operazioni per due argomenti diversi si devono fare più query. Esten-

dendo il SQL si può, ad esempio, effettuare più operazioni leggendo una sola volta la tabella, ed effettuando il minor numero di join possibile.

Analisi OLAP I tipi di operazione sono:

- roll up: riducendo il livello di dettaglio del dato, ovvero eliminare una o più clausole della groupby o navigare la gerarchia verso l'esterno;
- drill down: si aumenta il livello di dettaglio oppure si aggiunge una dimensione di analisi;
- slice and dice: consentono di ridurre il volume dei dati selezionando un sottogruppo dei dati di partenza;
- tabelle pivot: come viene mostrato il dato;
- ordinamento:

Queste operazioni possono essere fatte con più o una query.

Finestra di calcolo Una finestra di calcolo fa dei calcoli a partire da una query sottostante, la finestra ha 3 operazioni sottostanti:

- partizionamento (**partition by**): partizionamnto dei dati, divide i record in gruppi a partire dall'attributo selezionato;
- ordinamento (**order by**): si definisce il criterio di ordinamento delle righe all'interno dei partizionamenti;
- finestra di aggregazione (**over**): porzione di dati, specifica per ogni riga di dato, su cui effettuare dei calcoli;

Example 2.1

Data la tabella Vendite(Città, Mese, Importo), calcolare per ogni città la media delle vendite per il mese corrente ed i due precedenti.

```
1    SELECT Citta, Mese, Importo,
2        AVG(Importo) OVER (PARTITION BY Citta)
3                           ORDER BY Mese
4                           ROWS 2 PRECEDING)
5        AS MediaMobile
6    FROM Vendite
7
```

Quando la finestra è incompleta il calcolo è effettuato sulla parte presente, è possibile specificare che se la riga non è presente il risultato deve essere NULL.

Si può definire un intervallo fisico, superiore o inferiore.

```
    ROWS BETWEEN 1 PRECEDING AND 1 FOLLOWING
```

È possibile definire la tupla currente e quella che la precedono e che la seguono

```
    ROWS UNBOUNDED PRECEDING (o FOLLOWING)
```

Il rggruppamento fisico è specifico per quando i dati non hanno delle interruzioni.
Per definire un intervallo logico si utilizza il costrutto **range**.

```
    SELECT Citta, Mese, Importo,
        Importo DIV SUM(Importo) OVER (*) AS Boh,
        Importo DIV SUM(Importo) OVER (PARTITION BY Citta) AS Boh2,
        Importo DIV SUM(Importo) OVER (PARTITION BY Mese) AS Boh3
    FROM Vendite
```