Sistemi Informativi III – Riassunto dai lucidi

INFORMATION RETRIEVAL:

- Documento (d): unità di informazione reperibile espressa in formato libero;
- Archivio (D): insieme di documenti reperibili da un IRS;
- Necessità informativa (q): query;
- Rilevanza (RSV Retrieval Status Value): pertinenza, utilità rispetto alla query;
- Modalità di interazione:
  - Retrieval (intenzionale);
  - Browsing (non intenzionale);

Nell'IR testuale gli indici possono essere: parole, radici, frasi, parole o frasi da un vocabolario controllato, metadati (in aggiunta);

Misura di efficacia del Retrieval:

- $Precisione = \frac{|Rilevanti \wedge Reperiti|}{|Reperiti|}$;
- $Richiamo = \frac{|Rilevanti \wedge Reperiti|}{|Rilevanti|}$;
- Se si aumenta la recall si ottiene una diminuzione di precisione.

Generazione di un archivio di documenti testuali:

- è eseguita offline;
- consiste di 2 attività:
  - Localizzazione (inserimento);
  - Indicizzazione e generazione di struttura dati opportuna;

Indicizzazione:

- Finalità: rappresentare il contenuto del documento tenendo presente : esaustività e specificità;
- Tipologie: manuale (soggettiva), semiautomatica, automatica (oggettiva);
- Modalità:
  - estrazione diretta dal documento intero (full-text);
  - mediante fonti esterne (dizionari controllati);
  - tecniche associative (tesauri, pseudo-tesauri, clustering);
- Linguaggio di indicizzazione: linguaggio utilizzato per descrivere documenti e query;

Indicizzazione automatica di documenti testuali: consiste nell'estrarre dei termini indice che rappresentano in modo sintetico il contenuto del documento e utilizzare questi per l'indicizzazione. E' necessario fare una scelta bilanciata tra richiamo e precisione in quanto questi elementi influenzano pesantemente i risultati ottenuti.
Nella fase di indicizzazione devono essere compiute diverse operazioni:

- Analisi lessicale e selezione delle parole;
- Rimozione delle stopwords;
- Riduzione delle parole originali alle rispettive radici semantiche;
- Eventuale pesatura degli elementi dell'indice (significatività);

- Creazione dell'indice;

Tesauro associativo:

- è un grafo di parole, i cui nodi rappresentano parole e i rami rappresentano una (generica) relazione di similarità semantica tra le due parole.
- Vantaggi:
  - possono essere costruiti in modo completamente automatico a partire da una collezione di documenti;

Cluster di documenti: Raggruppamento di documenti simili in classi;
Strategie di clustering:
- globale: raggruppa i documenti basandosi sulle occorrenze degli indici nell'intera collezione;
- locale: raggruppa i documenti sulla base di un contesto definito dalla query (web);

La pesatura all'interno dell'indice può essere binaria se si considera solo la presenza/assenza del termine nel documento oppure in base alla frequenza del termine.

Frequenza e Rank:

- f(w) è la frequenza con cui w compare nella collezione;
- r(w) è l'indice di ranking (posizione) di w nella lista ordinata in funzione decrescente di frequenza, es: la parola più comune ha indice di rank uguale a 1;

Legge di Zipf:

Se le parole w di una collezione vengono ordinate r(w) in ordine decrescente di frequenza f(w) soddisfano la seguente relazione: $r(w) * f(w) = c$;
Collezioni differenti hanno costanti c diverse.
Nei testi in lingua inglese c tende a circa n/10 ove n è il numero di parole nella collezione.

Proposta di Luhn:

Viene proposto che la frequenza dell'occorrenza delle parole in un articolo fornisce un'utile misura dell'importanza del termine. E' altresì proposto che la posizione relativa all'interno di una frase delle parole aventi valori dati di importanza fornisce un'utile misura per determinare il significato delle frasi. Il fattore di significato di una frase sarà quindi basato sulla combinazione di queste due misure.

Criteri di indicizzazione basati sull'analisi di Luhn:

Pesatura dei termini indice: le parole più frequenti assumono un peso di significatività minore;
Stop lists: le parole più frequenti vengono eliminate dagli indici;
Parole significative: le parole più frequenti e meno frequenti vengono eliminate dagli indici;

Significatività dei termini indice:

$w_{td} = f_{td} * Discr - value_t$

f_{td} : frequenza del termine t in d è in relazione all'esaustività, fattore di recall;

$Discr - value_t$ è in relazione alla specificità, fattore di precisione;

Inverse document frequency:

$$idf_j = \log\left(\frac{N}{df_i}\right)$$

df_i : (frequenza del termine t_j nei documenti) è il numero di documenti in cui t_j appare

N : numero di documenti nella collezione

- favorisce la precisione;
- è alta se il termine appare in pochi documenti della collezione;

Significatività dei termini indice:

$$w_{ij} = tf_{ij} * \log\left(\frac{N}{df_i}\right)$$

Dopo aver eliminato le parole funzionali si calcola w_{ij} per ogni termine t_i in ogni documento d_j ;

Si assegnano ai documenti della collezione tutti i termini con valori alti di w_{ij} ;

poiché la frequenza assoluta del termine aumenta all'aumentare delle dimensioni del documento, è necessario normalizzare: $w_{ij} = \frac{tf_{ij}}{max\,tf_j} * \log\left(\frac{N}{df_i}\right)$

$max\,tf_j$ è la frequenza massima dei termini nel documento d_j

CREAZIONE DALLA STRUTTURA DATI DEGLI INDICI:

Scopi:

- velocizzare la ricerca di informazioni in collezioni semistatiche;
- l'indice viene costruito off-line prima di permettere la ricerca;

Indice a file inverted:

Scopi:

- velocizzare la valutazione delle query più frequenti
- la cancellazione di termini e dei documenti è rara;
- l'aggiornamento dei documenti è occasionale;

Per ogni termine indice viene indicato:

- la frequenza nella collezione;
- la lista di documenti che lo contengono;
- per ogni occorrenza del termine indice in un documento:
  - la frequenza del termine nel documento (tf);
  - la posizione nel documento (opzionale);

Algoritmo:

```
Per ogni documento d nella collezione D
    Per ogni termine indice t del documento d
        trova il termine t nel dizionario
        se t esiste aggiunge un nodo alla sua lista nel posting file
        altrimenti
            aggiunge t al dizionario
            crea un nodo nel file posting
```

al termine del processamento di tutti i documenti scrive il file su disco.

Indice lineare:

Il testo del documento viene diviso in blocchi e le occorrenze puntano ai blocchi ove compare la parola.

Vantaggi:
- numero di bit del puntatore ridotto rispetto a quello che punta all'esatta posizione del termine;
- tutte le occorrenze di una parola in un blocco hanno un solo riferimento;

Svantaggi:
- le ricerche per contesto richiedono più operazioni;

Indice a B-Tree:

Vantaggi:
- può essere acceduto velocemente (numero massimo di accessi di ordine O(log_d n) con n numero di livelli);
- conveniente per l'aggiornamento;
- semplice aggiunta di nuovi termini;
- economico utilizzo dello spazio in memoria;

Svantaggi:
- inefficiente per la ricerca sequenziale;
- tende a diventare sbilanciato dopo diversi aggiornamenti;
- se l'indice è su disco ogni nodo può richiedere un accesso al disco;

Indice a Suffix-tree:

Utilità:
- database genetici;
- ricerche complesse ad esempio per frasi e ricerca sequenziale;

Nozioni di base:
- tratta il testo di un documento come un'unica stringa;
- ogni posizione nel testo fino alla fine del testo è un suffisso;
- ogni suffisso è univocamente identificato dalla posizione iniziale dove appare nel documento;
- dal testo vengono selezionati dei punti indice che possono essere reperiti;
- i puntatori ai suffissi sono memorizzati nelle foglie dell'albero (suffix-tree);
- le parti comuni ai suffissi sono memorizzate solo una volta;

Vantaggi:
- adatto a ricerche complesse ad esempio per frasi e ricerca sequenziale;

Svantaggi:
- processo di costruzione costoso;
- dimensione dello spazio su disco dal 120% al 240% della dimensione della collezione;

Indice a signature:

Usa hash tables invece dei file inverted.
Idea:
- suddividere il documento in blocchi di lunghezza fissa;
- assegnare ogni blocco a una signature che viene usata per ricercare il documento effettuando un matching con la signature che viene associata alla query
- la signature associata al blocco si ottiene attraverso OR delle stringhe di bit associate a ciascun termine nel blocco da una funzione hash.

Vantaggi:
- bassa occupazione di memoria (dal 10% al 20% della dimensione della collezione) adatta per testi non troppo grossi;
- adatto per indicizzare dati multimediali;

Svantaggi:
- ricerca sequenziale sulle signature;
- eventualità di false drop;
- l'indice a file inverted è migliore nella maggior parte delle applicazioni rispetto all'indice a signature;

Ricerca sequenziale:

Se non c'è un indice o c'è un indice a blocchi è necessario procedere con la ricerca sequenziale.
Ci sono diversi algoritmi sia per la ricerca esatta che per quella simile per tenere conto dei possibili errori di stampa nella formulazione della query.

Ricerca Brute-Force:

Si confronta il pattern P con le stringhe in T che iniziano in tutte le possibili posizioni; alcune varianti usano una finestra sul testo.

Ricerca Knuth-Morris-Pratt

Si utilizzano le informazioni sul confronto nella finestra precedente per riposizionare la finestra ottimizzando la ricerca.

Misure di similarità tra stringhe:

Indicano la lunghezza della sequenza più lunga di caratteri comuni alle due stringhe.

Distanza di Hamming: numero minimo di caratteri che devono essere cambiati per trasformare una stringa nell'altra.
Distanza Edit o di Levenshtein: generalizzazione della distanza di Hamming. Stima il numero di trasformazioni (inserimenti, modifiche, cancellazioni) che devono essere effettuati per trasformare

una stringa nell'altra.
Espressioni regolari

Query per sistemi di IR basati su modelli di base
Tipologie:
- Linguaggi per utenti:
  - query booleane;
  - query vettoriali;
  - query booleane pesate;
  - query contestuali;
  - query sulla struttura dei documenti;
- Protocolli:
  - SFQL;

Valutazione di frasi: poiché durante l'indicizzazione vengono rimosse le stopwords e viene effettuato lo stemming del documento, una frase può corrispondere a più frasi nel documento originale.

Information Retrieval Distribuito:

- **Parallelo**: ha come obiettivo la minimizzazione dei tempi di esecuzione della query. Si realizza partizionando l'insieme dei documenti.
- **Distribuito**: sfrutta l'azione di più motori distinti. Le informazioni dei diversi motori non sono però confrontabili.

Calcolo del valore di Fitness della sorgente informativa:
1) documento prototipale: il valore di fitness è l'RSV del documento prototipale rispetto alla query
2) query-tipo definite da regole: il valore di fitness è calcolato tramite un insieme di regole antecedente => conseguente ove antecedente specifica gli argomenti delle query-tipo e il conseguente il valore di fitness;
3) query prototipale: il valore di fitness è calcolato valutando il grado di inclusione della query nella query prototipale;

Ricerca sul Web:

Motori di ricerca: permettono all'utente di formulare una query e restituiscono i documenti pertinenti;
Metamotori di ricerca: fondono insieme i risultati di diversi motori di ricerca per aumentare la pertinenza dei risultati;

Modelli di base di IR:

Il ranking è l'ordinamento dei documenti reperiti che riflette la rilevanza dei documenti rispetto alla query; è basato su assunzioni implicite nel modello del sistema di IR.

Modello booleano:

- ogni query booleana può essere riscritta in forma normale disgiuntiva;
- il retrieval è basato su un criterio decisionale binario (confronto esatto);
- non è in grado di produrre un ordinamento dei risultati;
- non si può controllare il risultato: produce come risultato o troppi o nessun documento;

Modello vettoriale:
- ha come scopo l'ordinamento dei risultati in funzione decrescente di rilevanza rispetto alla query;
- è basato sull'algebra lineare e rappresenta sia i documenti sia le query in uno spazio vettoriale n-dimesionale ove n è il numero di termini indice;

Modello vettoriale di G. Salton:
- la rilevanza è un concetto graduale ed è proporzionale alla similarità tra il vettore che identifica un documento e il vettore che identifica la query;
- indipendenza reciproca dei termini: la presenza contemporanea di coppie o di più termini nei documenti non è correlata in alcun modo;

Modello vettoriale di IR:
- rappresentazione basata su termini indice;
- T termini distinti sono usati per caratterizzare il documento;
- Ogni termine i è identificato da un vettore base dello spazio;
- t vettori sono linearmente indipendenti e formano una base ortonormale per lo spazio t dimensionale;
- ogni vettore nello spazio è una combinazione lineare di t vettori termini;
- l'r-esimo documento d_r può essere rappresentato come un vettore documento:

$$\vec{d}_r = \sum_{i=1}^{t} w_{ir} \vec{t}_i \quad ;$$

Prodotto fra vettori: $x \cdot y = |x||y| \cos\alpha$
similarità tra vettori $\cos\alpha = (x \cdot y)/|x||y|$

Misure di similarità:

- vettore documento $\vec{d}_r = \sum_{i=1}^{t} w_{ir} \vec{t}_i$;
- vettore query: $\vec{q}_s = \sum_{j=1}^{t} q_{js} \vec{t}_j$;
- $\vec{d}_r \cdot \vec{q}_s = \sum_{i,j=1}^{t} w_{ir} q_{js} \vec{t}_i \cdot \vec{t}_j$;
- $\mathrm{sim}(d_j, q) = \frac{\vec{d}_j \cdot \vec{q}}{|\vec{d}_j| \times |\vec{q}|} = \frac{\sum_{i=1}^{t} w_{i,j} \times w_{i,q}}{\sqrt{\sum_{i=1}^{t} w_{i,j}^2} \times \sqrt{\sum_{j=1}^{t} w_{i,q}^2}}$;

Algoritmo per il modello vettoriale:

Assunzione:
- t.idf fornisce l'idf di un termine t;
- q.tf fornisce la tf di un termine della query;

Begin

```
Score[] <= 0
For each term t in query q
        estrai posting list l
        For each entry p in l
                Score[p.docid] = Score[p.docid] + (p.tf * t.idf)(q.tf * t.idf)
```

I pesi w_{ij} e w_{iq} possono essere calcolati usando la frequenza assoluta o meglio ancora quella normalizzata per non privilegiare i documenti lunghi.

$$w_{i,j} = f_{i,j} \times \log\left(\frac{N}{n_i}\right)$$

$$f_{i,j} = \frac{freq_{i,j}}{max_l\, freq_{l,j}} \qquad idf_i = \log\left(\frac{N}{n_i}\right)$$

Il modello di IR vettoriale:

Vantaggi:
- la pesatura dei termini migliora la qualità del risultato;
- il confronto parziale permette di reperire documenti che approssimano le condizioni espresse nella query;
- la formula del coseno permette di ordinare i documenti in funzione del grado di similarità alla query;

Svantaggi:
- L'assunzione dell'indipendenza tra i termini non ha fondamento;
- i termini che non sono presenti nelle query non dovrebbero avere influenza sul retrieval;
- il linguaggio di query non è abbastanza espressivo;

Il modello probabilistico di Retrieval:

Obiettivo: inquadramento in ambito probabilistico del problema di IR;
La rilevanza: è un concetto binario (il documento è rilevante o non rilevante). Si stima la probabilità di rilevanza dei documenti;
Assunzione: data una query utente, c'è un insieme ideale rilevante che la soddisfa (R), la query è una descrizione delle proprietà di questo insieme ideale.
Ipotesi iniziale: la presenza nel documento di un termine della query è una indicazione a favore della rilevanza del documento;
Apprendimento: l'ipotesi iniziale viene rafforzata dalla continua osservazione della presenza del termine nei documenti rilevanti all'utente.
Vantaggi:
- i documenti sono ordinati in ordine decrescente di probabilità di rilevanza;
- incorpora un meccanismo di feedback di rilevanza;

Svantaggi:
- necessità di stimare le probabilità a priori $P(t_i|R)$;
- il metodo non tiene in considerazione i fattori tf e idf;

Logica Fuzzy:

E' un sovrainsieme della logica booleana introdotta per rappresentare concetti che stanno tra il completamente vero e il completamente falso.
In logica fuzzy il principio di non contraddizione e del terzo escluso non sono soddisfatti:

- non vale $A \cap \neg A = \emptyset$;
- non vale $A \cup \neg A = X$:
- non vale $A \cap A = A$;
- non vale $A \cup A = A$;

Operatori OWA:

$$W = [w_1, w_2, \ldots, w_n] \quad \text{con} \quad w \in [0,1] \quad \text{e} \quad \sum_i w_i = 1$$

$$F(a_1, \ldots, a_n) = [w_1, \ldots, w_n] \begin{bmatrix} b_1 \\ \ldots \\ b_n \end{bmatrix} = \sum_{i=1}^{n} w_i * b_i$$

$$orness(W) = \frac{1}{(n-1)} \sum_{i=1}^{n} w_i (n-i)$$

Linguaggio con operatori linguistici di aggregazione:

Obiettivo:

- superare l'aggregazione “rigida” dei termini;
- evitare l'uso di operatori la cui semantica è spesso ambigua;
- semplificare la formulazione di query senza perdere in espressività;

Soluzione:

- definire quantificatori linguistici per aggregare i termini nelle query (tutti, la maggior parte, almeno k...)
- definizione dell'operatore e possibilmente per specificare vincoli opzionali;

Modello di IR Latent Semantic Indexing:

I modelli classici di IT producono risultati scandenti:

- tra i documenti reperiti ci possono essere documenti non in relazione con gli altri;
- i documenti rilevanti che non contengono termini indice non vengono reperiti;
- la correlazione tra termini e documenti basata su criteri statistici è inaffidabile:
  - termini differenti possono avere lo stesso significato;
  - lo stesso termine uò avere significati differenti a seconda del contesto;

l'obbiettivo è quello di trovare una modellazione più affidabile della correlazione tra termini e documenti.
Consiste nell'utilizzare un insieme ristretto di termini invece di usare tutti quelli presenti, però questo riduce la precisione, è quindi necessario trovare un giusto equilibrio tra numero di termini e precisione.

Conclusioni:

- migliora del 20-30% il retrieval basato su termini indice;
- metodo completamente automatico e ampiamente applicabile;

Information retrieval associativo:

- i meccanismi associativi di IR sono basati sull'utilizzo di associazioni per migliorare i risultati del retrieval;
- permettono di estendere l'insieme dei documenti reperiti da un sistema di IR con documenti non direttamente indicizzati dai termini specificati nelle query;
- sono basati sul concetto di indice associativo, un dizionario che associa descrittori ad altri descrittori mediante relazioni termine-termine e documento-documento;

Tesauri

Un tesauro per IR è un indice associativo in cui i descrittori sono termini:

- broader terms (termini più generali);
- narrower terms (termini più specifici);
- related terms (termini sinonimi);
- used for (termini in uso);

I tesauri però non esistono per tutte le lingue, per questo si usano anche gli pseudo-tesauri, generati statisticamente da una collezione di documenti oppure i fuzzy tesauri;

Relevance Feedback

Viene utilizzato per aumentare la precisione delle query e sfrutta le preferenze dell'utente ritornate dopo l'effettuazione della query per rilevare quali documenti erano effettivamente rilevanti per esso.

Difetti: scarsa diffusione;

- gli utenti non sono inclini a fornire indicazioni esplicite;
- spesso le query modificate sono complesse e richiedono più calcoli;
- diventa complicato per un utente capire perché un documento è stato reperito;

Ricerca sul web

I motori di ricerca servono per la ricerca di informazioni e utilizzano come archivio di documenti la parte pubblica del web.

I search engine sono composti da:

- web crawlers;
- modulo di indexing;
- modulo di retrieval;

I web crawlers sono gli agenti software incaricati di scandagliare il web per recuperare tutti i documenti che devono essere indicizzati dal motore di ricerca; ne esistono di diversi tipi con modalità di funzionamento differenti, alcuni sono open source altri sono proprietari.

Una volta recuperato il documento questo viene analizzato e posto all'interno dell'indice del SE per poter essere poi reperito dalle query dell'utente. La fase di indicizzazione è molto delicata e

complessa e deve tener conto della dinamicità del contenuto del web; molti siti infatti variano i propri contenuti molto spesso, anche più volte al giorno, e i SE devono essere in grado di analizzare più frequentemente le pagine che cambiano spesso.

Gli algoritmi di indicizzazione sono quasi sempre brevettati e a volte molto simili tra loro, per questo sono pochi i SE “originali”.

Le pagine reperite dalle query vengono ordinate per mezzo di algoritmi di ranking proprietari che combinano moltissimi fattori per determinare l'indice di importanza della pagina. Solitamente i criteri più importanti e utilizzati da tutti i motori di ricerca sono il numero di inlink e di outlink della pagina nonché la loro qualità.

I sistemi di IR vengono valutati per giudicare il livello di precisione e di attendibilità, per questo vengono effettuate prove con delle collezioni di documenti e delle query prestabilite e studiate per valutare le prestazioni del sistema. Queste vengono giudicate in base alla rilevanza e alla posizione dei documenti reperiti tra i risultati.

Geographic Information Systems (GIS):

Cosa sono:

Sono sistemi che permettono di rappresentare la spazialità delle informazioni e sono utilizzati in moltissimi campi.

Per rappresentare oggetti spaziali è necessario tenere in considerazione le caratteristiche che questi possono avere; un oggetto nello spazio infatti può avere:

- confini netti e ben definiti;
- confini sfumati e non ben definiti;
- proprietà che variano in modo continuo;
- proprietà discrete;

Esistono poi diversi modi di rappresentazione degli oggetti, i principali sono:

- rappresentazione vettoriale;
- rappresentazione raster;

Entrambi hanno vantaggi e svantaggi.

Per rappresentare oggetti in uno spazio geografico è poi possibile utilizzare differenti sistemi di coordinate, i principali sono:

- coordinate cartesiane;
- coordinate polari;

Esistono poi diversi sistemi di mappatura:

- mappatura piana;
- mappatura conica;
- mappatura cilindrica;

La cartografia ufficiale italiana utilizza il sistema Gauss-Boaga (dal 1940) con due diversi fusi:

- fuso ovest e fuso est coincidenti con i fusi 32 e 33 UTM (9° e 15° long est)
- punto di emanazione (coordinate di Monte Mario a Roma)

Tipi di entità spaziali:

- entità 0-dimensionali: punti;
  - utili per rappresentare oggetti non ben definiti o con dimensioni piccole in relazione alla scala;
- entità 1-dimensionali: linee;
  - usate per la rappresentazione di strade, corsi d'acqua e confini;
- entità 2-dimensionali: poligoni;
  - per rappresentare edifici, regioni, ...;

Rappresentazione a spaghetti:

Rappresenta la geometria di una collezione di oggetti spaziali senza esplicitare le relazioni topologiche, ma considerando ogni oggetto indipendentemente dagli altri; le primitive sono i punti, le linee, le poli-linee, i poligoni e le regioni.

Caratteristiche:

- ridondanza: spigoli in comune tra oggetti sono ripetuti;
- semplicità;
- facilità di aggiornamento;
- inconsistenza: il confine tra due oggetti può essere rappresentato da due poli-linee diverse non coincidenti;
- mancanza di esplicitazione delle relazioni topologiche;

Rappresentazione a rete:

Utile per rappresentare applicazioni 1D modellate con grafi quali le reti di servizi (idriche, stradali, telematiche, ecc.).
Rappresenta esplicitamente le relazioni topologiche tra punti e polilinee.
Si introducono oltre ai punti e alle polilinee:

- nodi: un nodo è un punto che connette una serie di archi;
- spigoli: uno spigolo è una polilinea che inizia in un nodo e termina in un altro nodo definendo un possibile cammino tra un insieme di punti;
- una rete può essere piana (ogni intersezione è un nodo) oppure non piana (ammette tunnel e ponti);

Vantaggi:

L'esplicitazione della relazione di connessione tra nodi rende tale rappresentazione adatta ad applicazioni in cui sia necessario ricercare il miglior percorso in un grafo.

Svantaggi:

Non vi è possibilità di rappresentare le relazioni tra oggetti 2D.

Rappresentazione topologica

Utile per rappresentare applicazioni 2D in cui si faccia un uso intensivo di analisi della topologia degli oggetti geografici.

E' una rappresentazione che introduce una suddivisione in poligoni adiacenti dello spazio 2D basata su due elementi di base:

- nodi: un nodo è un punto che connette una serie di spigoli;
- spigoli: uno spigolo è una polilinea che inizia e termina in un nodo e separa due poligoni.

Vantaggi:

- l'esplicitazione della relazione di connessione tra poligoni la rende adatta alla valutazione di query topologiche;
- poiché i poligoni condividono gli spigoli, è più semplice l'aggiornamento (non si rischia di introdurre inconsistenze)

Svantaggi:

- alcuni poligoni possono non avere una semantica reale;
- la mappatura è costosa poiché richiede spesso la cancellazione di spigoli tra poligoni che rappresentano la stessa entità geografica.

Database geografico:

Tema: informazione georeferenziata relativa a un certo argomento (es.: città, fiumi, dati epidemiologici)

Oggetto geografico: è un'entità del mondo reale che ha due componenti:

- una descrizione: insieme di attributi alfanumerici descrittivi o tematismi;
- una composizione spaziale o georiferimento: attributo geometrico che descrive e definisce la geometria dell'oggetto geografico e la topologia dell'oggetto geografico.

Un oggetto geografico può essere:

- atomico;
- complesso: costituito da oggetti geografici atomici;

Operazioni spaziali semplici:

- selezione di un tema;
- unione di temi;
- sovrapposizione di temi;

Operazioni di selezione geometrica:

- **windowing**: è un'operazione di overlay tra un tema e una finestra definita dall'utente sul dominio spaziale. Il risultato p un tema definito da tutti gli oggetti geografici la cui componente geometrica interseca la finestra;
- **puntamento**: simile al windowing; il risultato è un tema definito da tutti gli elementi geografici la cui componente geometrica contiene il punto;
- **clipping**: definisce la porzione di un tema localizzato in un'area specificata da una finestra. La componente geometrica del tema viene calcolata come l'intersezione tra la componente geometrica del tema e la finestra di clipping.

Operazioni spaziali:

- operazioni metriche: calcolano la proprietà metriche degli oggetti geografici di un tema;
- operazioni topologiche: calcolano proprietà topologiche degli oggetti geografici di un tema;
- operazioni di interpolazione / estrapolazione: stimano il valore di attributi descrittivi in punti o regioni del dominio spaziale a partire dal valore noto di tali attributi in punti sparsi;
- operazioni di localizzazione: dato un insieme finito di possibili locazioni di risorse sul territorio, scegliere la locazione in cui l'ammontare delle risorse superi/corrisponda a una certa quantità;
- operazioni di allocazione: dato un insieme di punti di richiesta di risorse e un insieme di punti di distribuzione, valutare come distribuire le risorse (applicazioni di pianificazione urbana);

Svantaggi del modello relazionale:

- violazione del principio di indipendenza dei dati: per formulare una query è necessario conoscere la strutturazione della componente spaziale;
- performance scadenti: sono necessarie diverse relazione per rappresentare la componente spaziale con conseguente inefficiente accesso;
- non sono disponibili operatori spaziali (geometrici e topologici);

Rappresentazione dei dati spaziali:

Per rappresentare i dati spaziali di cui un GIS necessita spesso vengono utilizzati i diffusissimi database relazionali, ma questi non sono progettati per la gestione di informazioni spaziali e per

questo è necessario estenderli per mezzo di opportuni tipi di dati spaziali astratti (Abstract Spatial Data Types – ASDT). Un esempio comune di database relazionale adattato è PostGIS, ovvero una collezione di funzioni e ASDT che estendono le funzionalità di PostgreSQL per adattarle alla gestione di dati spaziali.

Queste funzionalità aggiuntive comprendono sia i tipi di dati (punti, linee, poligoni,...) che le funzioni per la manipolazione (traslazione, rotazione,...) e il confronto e la misura (distanza, area, perimetro, punto-in-poligono, ...).

Rappresentazione digitale del territorio:

La rappresentazione digitale del terreno è molto utile in diversi campi, da quello dell'ingegneria civile a quello delle telecomunicazioni dove i modelli computerizzati permettono di avere una conoscenza del territorio senza dover effettuare sopralluoghi e misurazioni.

Per visualizzare i contenuti di un GIS è poi possibile effettuare un “rendering” delle informazioni in un formato raster (immagini JPEG, PNG, BMP...) o meglio esportarli in un formato vettoriale come SVG che permette di visualizzare a diverse fattori di ingrandimento senza perdere nel dettaglio.

Per lo scambio di dati esiste un formato standard chiamato GML (Geographic Markup Language) basato su XML e diversi formati de facto portati dai software proprietari più diffusi in questo ambito (i più famosi e praticamente lo standard de facto in generale sono i formati della ESRI).