

Basi di dati multimediali

Maria Luisa Sapino

(orario ricevimento: lunedì ore 14-16)

Modalità di Esame

- 5 appelli all'anno:
 - » Fine giugno 2011
 - » Meta' luglio 2011
 - » Meta' Settembre
 - » Fine Settembre
 - » Febbraio/marzo 2012

- L'esame e' orale.

AVVISO

- Dal 21 al 25 marzo parteciperò alla conferenza EDBT 2011 ad Uppsala.
- NON CI SARA' LEZIONE DI BDM

Preliminary Discussions

- What is media?
- What is multimedia?
- What is hypermedia?

Basi di dati multimediali

- Prerequisiti: corso di Basi di dati e Lab/Sperimentazione della laurea triennale
- Corsi correlati:
 - Fondamenti della Comunicazione
 - Architetture di basi di dati
- Materiale didattico:
 - Appunti
 - lucidi usati a lezione e draft del libro
- Candan, Sapino "Data Management for Multimedia Retrieval", Cambridge University Press, maggio 2010.

Sample application

- Police investigation...
 - Video data (surveillance cameras)
 - Audio data (telephone wiretaps)
 - Image data (surveillance, mugshots)
 - Document data (police reports)
 - Conventional data (bank records, employment records, police records)
 - Geographic data (maps)

Sample multimedia query

- “Find the records of every criminal who look like the person seen in “surv_im.gif” and who had a bank transfer of more than \$500,000 within the last 5 months. Return all police reports which mentions such persons and their past accomplices.”

Interactivity

- 100ms interaction deadline
 - resource allocation
 - pre-fetching/caching
- Subjectivity and personalization of content
- Interaction structure

Semantic Heterogeneity

- Spatio-temporal-hierarchical dimensions
 - modeling
 - specification
 - indexing,
 - retrieval, and
 - visualization methods
- User- and context-dependence, subjectivity
- Availability at various quality levels

What is a data model?

Physical Heterogeneity

- Volume
 - storage,
 - delivery, and
 - processing
- Quality/cost trade-off
 - increases robustness, graceful degradation

What is a database management system?

- A system which allows access to a collection of data
 - User specifies **what** to see
 - System retrieves the corresponding data from the collection
 - The system presents the retrieved info to the user
- ..different from **browsing**

Queries

- Metadata queries
- Example queries
 - Exact
 - Partial match
- Object queries
 - visual similarity
 - semantic similarity
 - spatial similarity

Why “image” database?

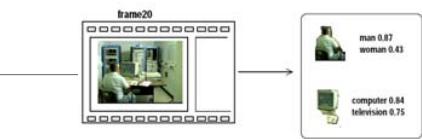
- Size of data
- Properties of data
 - Visual: image processing
 - Semantic: pattern recognition
- Similarity-based retrieval
 - Similarity-based query processing
 - new index structures
 - relevance ordering
- Query language
 - How to let users specify what they want?

What is an image?

- 2D matrix of values
- Collection of objects and their spatial relationships
 - An object is an entity within an image
 - visual
 - semantical

What kind of images?

- Mug shots, cat-scans (cat=Computerized Axial Tomography), fingerprints
- News, advertisement, family photos
- Surveillance
- Video frames



What is an image database?

- A collection of images
 - local or web
- A query processor (indices etc.) which
 - maps user query into data model
 - retrieves the relevant images
- An information visualization system which shows results to the user

What are the features of interest?

- Colors, color histograms
 - “sunny day”, “sea”
- Edges
 - “maps”, “aerial surveillance”
- Texture
- Image segments
 - shape, location, color
- Objects
 - visual features, semantics
- Metadata, captions

What kind of queries?

- Find me all images created by “John Smith”
- Find all images which look like “im_ex.gif”
 - Find me top-5 images which look like “im_ex.gif”
- Find all images which look like “sketch.bmp”
- Find all images which contain a part which looks like

Example

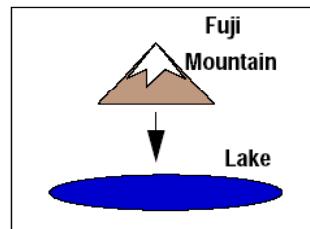
```

select image P, object object1, object object2
where P contains object1
and P contains object2
and object1.semantical_property s like "mountain"
and object1.image_property image_match "Fuji_mountain.gif"
and object2.semantical_property is "lake"
and object2.image_property image_match "lake.image.sample.gif"
and object1.position is .above object2.position
  
```

What kind of queries?

- Find all images of sunny days
 - advertisement
- Find all images which contain a car
- Find all images which contain a car and a man who looks like ”mugshot.bmp”
 - surveillance
- Find all image pairs which contain similar objects
 - data mining

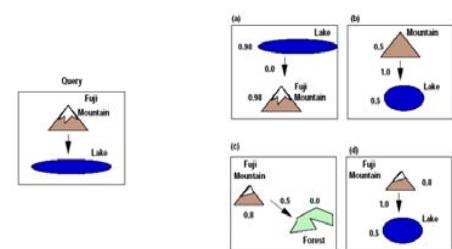
QBE (visual representation)



What kind of queries?

- Find all objects contained in images of sunny days
 - Find all images which contain two objects
 - first object looks like “im.gif”
 - second object is a car
 - first object is to the right of second obj.
- and return the semantics of these two objects.

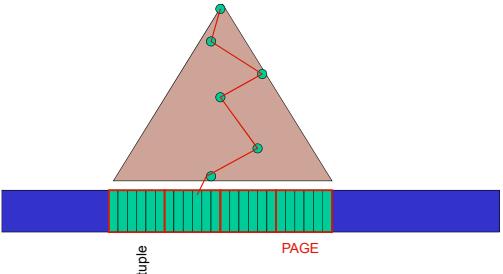
Query...and results...



Relational databases

- Data is
 - textual
 - numerical
- This is the main assumption for
 - storage
 - query processing
 - optimization

Index structures



Relational databases

- Information is in tabular form
 - Example: Information about an employee
- Schema describes the content
- A key uniquely identifies a given tuple
- Each attribute has a domain

Attribute	Schema		
NAME	SSN	OFFICE	DESC
..
J. Doe	555-5555	GWC 999	Asst. prof
J. Smith	333-3333	GWC 989	Prof
..

Algebra

- A set of data manipulation operators
- Relational algebra (operates on relations)
 - Select (σ)
 - Project (π)
 - Cartesian product (\times), join
 - Union (U)
 - Intersection (\cap)
 - Difference ($-$)
- Procedural (non-declarative)

Index structures

- Disk is divided into logical units, called “pages”
- A relation/table is stored contiguously
- Each page contains a certain number of tuples
- An index structure is created for rapid ($O(\log N)$) access to information

Calculus

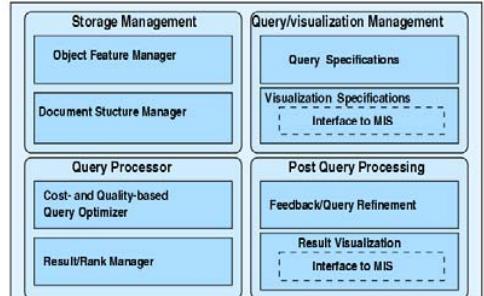
- A query language should be declarative:
 - Say what we want
 - query optimization
 - Don’t say how we get it
 - no optimization possible
- ```
{t.name | (t ∈ Employee) and (t.salary < 1000) and
(exists t2 (t2 ∈ Students) and (t2.grade-average > 3.7)
and (t.ssn = t2.ssn)
)}
```

## SQL

- Based on relational calculus
- ```
select <attribute_list>
from <relation_list>
where <condition>
```

```
select t.name
from employee t, student t2
where (t.salary < 1000) and
      (t2.grade-av> 3.7) and
      (t.ssn = t2.ssn)
```

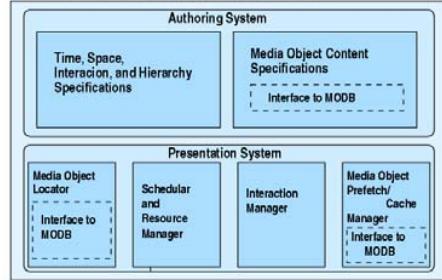
Multimedia Object/Document Base (MODB)



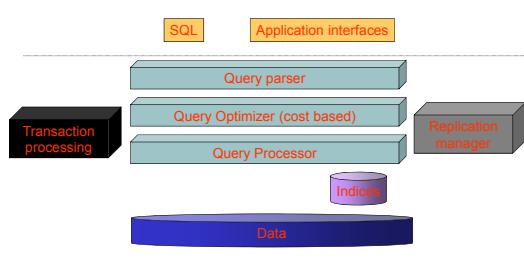
Relational databases

- Business applications
- Data model is relational
- Queries are exact/declarative
- Updates are important
- Concurrency is important

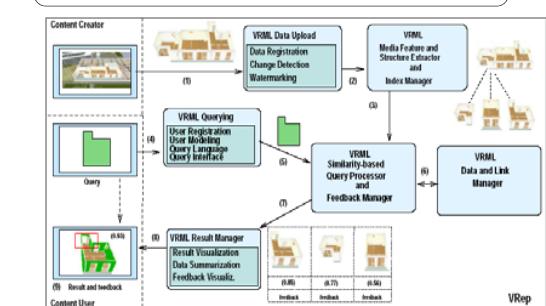
Multimedia-object Integration System (MIS)



How does a database look like?



Example: X3D/VRML Archive



Relational databases (??)

- Business applications
- Data model is relational
- Queries are exact/declarative
- Updates are important
- Concurrency is important

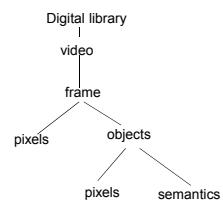
Other problems?

- It does not capture the semantical structure of the data well
- Hierarchies:
 - Aggregation hierarchy
 - Inheritance hierarchy

Shortcomings...

- Image data doesn't fit into tuples
 - Media data need to be kept separately
- No image comparison
- No partial match processing
- No ranking
- Not computationally complete
 - Media processing requires more computational power.

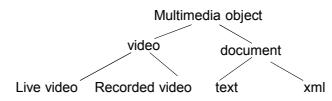
Aggregation hierarchy



Solutions

- Use a host language and embed database queries in it (**relational approach**)
- Provide more computational power in the data model itself (**object-oriented approach**)

Inheritance hierarchy



OODB

- Object oriented databases provide
 - Higher computational power
 - Aggregation hierarchies
 - Inheritance hierarchies
- They model the real world better!
 - Everything is an object
- You can define your own external methods

E.image_similar_to (c.image)

Object Relational Databases

- Benefits from both
 - Relational technology
 - tuples
 - SQL
 - Object technology
 - User defined functions
 - User defined abstract data types (ADTs)

OODB

- Business applications
- Multimedia (??)
- Data model is object oriented
- Queries are exact
- Queries are procedural (some declarative languages)
- Concurrency/updates are important

Shortcomings...

- No partial match processing
- No ranking
- Query processing is cost driven
 - not “similarity” driven

Shortcomings...

- Too much overhead
 - Optimization is hard
- No partial match processing
- No ranking
- Query processing is cost driven
 - not “similarity” driven

What else?

- Deductive databases
 - Logic based
 - Boolean queries
- Fuzzy databases
 - usually logic-based, but not boolean
 - nothing is *true* or *false*
 - results are not-exact (like multimedia queries)

What else?

- Spatial/Temporal Databases
 - Scientific, geographic applications
 - Data model is vector or interval based
 - Queries
 - Range queries
 - Nearest neighbor queries
 - Queries are declarative or visual

...and

- Image databases
 - Data model is feature-vector based
 - Multiple features
 - Color
 - Texture
 - Each feature represented as a vector space
 - Structure may or may not be available
 - Queries
 - Query-by-example
 - Ranking
 - Feedback (user-to-system, system-to-user)

What else?

- Data mining
 - Business, scientific applications
 - Relational data model
 - Queries: find
 - patterns,
 - rules,
 - classes, or
 - outliers

Research Issues

- Data model
 - Content, features of interest
 - Information extraction/integration
- Query Language
 - matches the data model
 - captures user's interest

What else?

- Semi-structured data management
 - Most data has a well-defined structure (schema)
 - In SSD, there is no common schema
 - each object describes itself
- Queries
 - Structure-based

Research Issues

- Query processing
 - Online vs. off-line information extraction
 - Indices for different media
 - Optimization of queries with different media
 - Similarity-based retrieval, ranking
 - Relevance feedback

Research issues

- Storage/delivery
 - How to store data in different formats
 - How to retrieve data efficiently
 - speed
 - precision and recall
 - How to transmit large, continuous data (video)
- Visualization
 - How to visualize/present results of a query which may contain multiple types of data (images, video, audio)

query su dbm:
trova tutti i records su un certo criminale che assomiglia a persona in immagine img.gif e che ha ricevuto bonifico > 500mila dollari"
restituisce tutti i rapporti di polizia che menzionano questa persona

*riconoscimento
*accesso ai dati bancari

gestione di dati multimediali: l'aspetto fondamentale è ETEROGENEITÀ dei dati che si presenta sia a livello sintattico che semantico

eterogeneità semantica si manifesta nelle dimensioni spazio-temporali:
perché gestire lo spazio? perché parlando di immagini ci interessano relazioni spaziali di oggetti nell'immagine, quindi bisogna sapere descrivere lo spazio -> ci interessa per fare query su punto dell'img in cui un oggetto si trova, località dell'oggetto individuato
perché gestire tempo? --> video, cosa succede prima/dopo, es: arma che compare prima o dopo omicidio ci dice se è stata usata o meno
sempre possibile partizionare spazio in parti più piccole, quindi diversi gradi di granularità; stessa cosa per il tempo (secondi, minuti, ore, ecc.); devo conciliarli per fare delle query e gestire le eterogeneità di rappresentazione

per ciascuna dimensione spazio-temporale bisogna decidere come modellarle, cioè astrarre le caratteristiche di interesse, e rappresentarle opportunamente: bisogna trovare forme analoghe a schemi ER, cioè forme di astrazione di FEATURES (caratteristiche) che caratterizzano gli oggetti, quelle che scartiamo non possono essere usate nella query e quindi non concorrono al risultato finale

- * bisogna studiare diverse forme di modellazione: es: query come la faccio? testuale? passo immagine da confrontare? passo uno sketch con le caratteristiche salienti? come specifico features di interesse?
- * indicizzazione: b-alberi in modello ER sono struttura indice tradizionalmente utilizzata. b-alberi presuppongono relazione di ordine sui dati: come ordino dati multimediali?
- * RETRIEVAL: restituire risultati di interesse
- * visualizzazione dei risultati in modo efficace per l'utente

INCERTEZZA dei risultati di una query (a differenza del modello ER): es. persona bionda? quale è il colore esatto? quando una persona smette di essere bionda e diventa castana? alto-basso, sotto-sopra, ecc... sono tutti concetti relativi. di solito i risultati sono ordinati in base al grado di soddisfacimento che plausibilmente corrisponde a quello dell'utente.

La SOGGETTIVITÀ e la dipendenza dal contesto dell'utente sono aspetti da tenere in conto: aiutano a dare un senso particolare ad un oggetto
* il contesto rende un certo dato importante o no; la stessa relazione di importanza potrebbe essere diversa in contesti diversi
* soggettività entra in gioco quando ci sono aspetti che non sono oggettivamente catturabili (es: espressione serena di una persona nell'img)

i nostri sistemi devono tenere conto anche di queste caratteristiche, con RELEVANCE FEEDBACK, ovvero l'utente dà un feedback sulla rilevanza degli oggetti che gli sono stati restituiti, fb elaborato dal sistema per correggere la query per tenere conto della rilevanza data dall'utente a certi aspetti piuttosto che ad altri.

altro aspetto importante (per img e video): disponibilità di dati di qualità, quindi risoluzione immagini e # frames di un video -> ovviamente la qualità incide! bisogna tenere conto anche di questo nel definire un meccanismo di retrieval nelle dbm, perché l'utente può anche decidere di volere una certa qualità in base alle sue esigenze

ETEROGENEITÀ A LIVELLO FISICO:

- * quando devo combinare testo+video+immagini, devo tenere conto che (es: tg sottotitolato), laddove un certo spazio di memoria mi sarebbe sufficiente per memorizzare i sottotitoli, quando ci metto anche il video ovviamente aumenta lo spazio richiesto! richiedono ordine di grandezza diverso, quindi sistema di indicizzazione e di retrieval e processamento devono anche tenere conto di queste eterogeneità a livello di occupazione di spazio, di software richiesti, di banda richiesta (es: ne serve di più per scaricare un video!)
- * tradeoff tra qualità e costo: massimizzare la qualità e minimizzare i costi, però sono obiettivi che vanno in contrasto quindi bisogna giocare sul bilanciamento tra queste due dimensioni, quindi cercare di migliorare la qualità senza far crescere troppo il costo; inoltre, man mano che la qualità degrada, vorrei che il degrado fosse una cosa graduale, quindi distribuire la qualità in modo da non fare sbalzi troppo bruschi di qualità (es: da qualità blueray a 5 secondi di static): per l'utente percepire un degrado più graduale può essere meglio, ma è soggettivo; il sistema cmq deve gestire anche questi aspetti di compensazione

DBM

- * interattività: un sistema soddisfacente per l'utente deve presentare i risultati in tempi brevi
- * prefetching e caching:

- prefetching -> recuperare dati dalla memoria in cui i dati sono localizzati in anticipo rispetto a quando mi serviranno
- caching: tenere in memoria oggetti già visti e che si pensa possano essere richiesti nuovamente

noi però ci concentreremo sulla GESTIONE delle dbm.

si vorrebbe anche personalizzare la presentazione di risultati in base alla soggettività dell'utente e al suo contesto, es: se cerco "aviaria" voglio dare risultati diversi a bambino/adulto/medico, a livello di qualità e contenuto
struttura di interazione: come l'utente può interagire col sistema

esigenza di modellare i dati multimediali eterogenei:
modello di rappresentazione di tutte e sole le caratteristiche importanti dei dati e delle operazioni sui dati: quali sono le features e cosa posso farne?

DBMS: sistema che mi consente di accedere a collezione di dati multimediali consentendo all'utente di specificare cosa vuole vedere, far sì che il sistema recuperi i dati rilevanti rispetto alla richiesta dell'utente e presenti i dati all'utente.
nel modello ER: dichiarazione di cosa si vuole vedere (in SQL), sistema accede ai dati e restituisce relazione che soddisfa la query. definiamo un analogo per le DBM!

fare il RETRIEVAL è diverso da BROWSING!
* browsing non sappiamo fin da subito dove vogliamo arrivare
* retrieval dichiariamo subito a cosa siamo interessati

=====

QUERIES di natura diversa: dobbiamo conciliarle
* query su metadati (dati che descrivono i dati): es. autore di una fotografia, la data, regista di un film, sono dati solitamente di carattere testuale che descrivono e completano le info su un dato -> si può fare query solo su metadato senza entrare nel merito della rappresentazione dell'oggetto stesso
* query basate su esempio: di tipo esatto o match parziale. "trovami tutti gli ogg. che assomigliano a questo", io dò l'oggetto che voglio trovare e voglio che vengano trovati tutti quelli simili.
- esatta: restituisco tutti e soli i risultati che corrispondono in tutto e per tutto alla query, è query binaria
- parziale: voglio che siano restituiti tutti gli oggetti che in qualche modo assomigliano all'oggetto dato e che i risultati mi siano presentati in ordine decrescente di somiglianza (come fa Google!)
* query su oggetti:
- somiglianza visuale
- somiglianza semantica
- somiglianza spaziale
es. foto di un bambino che sorride, poi foto dello stesso bambino che piange: visualmente si assomigliano (stessi capelli, magari stesso vestito), nei termini dei colori e delle forme sono simili, ma semanticamente sono abbastanza diverse.
bambino che sorride e sketch di uno smilie :), visualmente non si assomigliano, ma semanticamente sono vicini: sono due rappresentazioni diverse di un sorriso! i due aspetti non sempre vanno di pari passo: somiglianza visuale non sempre matcha somiglianza semantica:
bisogna trovare modo di esprimere la priorità della somiglianza che ci interessa
somiglianza spaziale: posizione reciproca di oggetti che si trovano nel dato, può essere più o meno rilevante a seconda del contesto (es: partita di calcio)

=====

faremo riferimento quasi sempre alle immagini: ci interessano le basi multimediali per introdurre tecniche che poi si possono applicare anche a casi molto diversi

importanza dei concetti di NATURA e PRECISIONE dell'informazione

nelle BDR tutti i dati sono esatti: i dati originari possono però essere stati raccolti con strumenti imperfetti (es: altezza di una persona) e quindi sono intrinsecamente imprecisi

confronti tra oggetti sono imprecisi, la somiglianza si misura in maniera imprecisa e dipende da quali criteri uso per misurare

immagini: sono più complesse del testo, inoltre un video è semplicemente una sequenza di immagini + eventualm. audio, quindi usiamo immagini come oggetto di studio e poi aggiungiamo la dimensione TEMPO per parlare dei video

2 approcci diversi alla definizione di immagine: quale è migliore? dipende!
* dal punto di vista della rappresentazione su display un img è semplicemente una matrice 2D di pixels: se io voglio fare interrogazione su colore dominante, io devo contare quanti pixel hanno un certo colore; se invece voglio fare interrogazioni a livello semantico? non basta la

matrice di pixel, è troppo di basso livello e direttamente misurabile e quantificabile (sono numeri!), è oggettiva
* se voglio fare query più complesse, bisogna dare caratterizzazione semantica in cui si apprenda quali sono gli oggetti rappresentati e quale sia la loro collocazione spaziale, e gli oggetti rappresentati sono a loro volta sotto-immagini che hanno caratteristiche visuali e semantiche

=====

per fare DB di immagini cosa ci serve?

- * DBM è una collezione di immagini, ma non solo
- * non può prescindere dalla collezione di immagini, ma su questa collezione definisce una ORGANIZZAZIONE, consente meccanismi di INTERROGAZIONE e mappatura di query utente nel modello dei dati per poterla confrontare con gli oggetti del DB (es. calcola matrice di pixels della query), fare confronti con meccanismi di confronto, retrieval di immagini rilevanti (quelle che SI PENSA meglio corrispondono alla query - non c'è concetto di ESATTO! - se non sui metadati :))
- * sistema di visualizzazione che presenti all'utente i risultati: è importante perché è quello che interessa all'utente!

=====

DBM deve consentire di gestire e rappresentare le proprietà dei dati (di carattere visuale o semantico)

- * es. colori, histogramma dei colori, tessitura che si ripete che permette di riconoscere materiali nell'immagine (es: cemento piuttosto che erba): caratteristiche che possono essere estratte tramite analisi dell'img e che memorizziamo
- * poi ci sono caratteristiche semantiche che vengono estratte tramite tecniche di riconoscimento (es. struttura più o meno costante dei volti) basate su pattern comuni (es: volto è ovale, ha due ovali più piccoli che sono gli occhi, ecc...)

=====

QUERIES

è necessario un modulo che consenta di fare queries basate su somiglianza ed evitare confronti inutili: qui è ancora più importante che nelle BDR perché i dati sono di grandi dimensioni!

presentare i dati ordinati in base alla loro rilevanza/somiglianza

serve linguaggio per specificare le query (tipo SQL)

=====

quali tipi di immagini?

foto segnaletiche, impronte digitali, TAC; frames all'interno di un video -> analizzati per riconoscimento di forme e ad ogni forma è assegnato con una certa probabilità un significato: come combinare questi valori? media aritmetica o altro? vedremo in seguito cosa comportano diverse scelte in merito...

=====

quali sono le features di interesse?

non riusciamo mai a rappresentare i dati nella loro completezza di tutti i suoi aspetti, quindi si ha sempre a che fare con dati parziali, ossia SURROGATI dei dati che catturano solo un certo numero di aspetti, di proprietà (features) dell'oggetto stesso

tutte le features che memorizzo sono quelle su cui potrò fare delle query, quindi non è una scelta banale!

tipicamente le features di interesse sono di 6 categorie principali:

- * colori -> istogramma dei colori: ad ogni colore associo un contatore che mi dice quanti pixel hanno quel colore, il num. di colori che memorizzo è il num di BINs (es 256 colori), posso fare query del tipo "quale colore dominante?" e su distribuzione dei colori
- * contorni -> evidenziano il punto di stacco di colore di una superficie che concorre a riconoscere forme in un'img
- * texture
- * segmenti di immagini che ha loro volta hanno caratteristiche di forma, localizzazione, colori
- * oggetti nell'immagine, che forma hanno, dove si trovano
- * metadati (dati che descrivono i dati)

=====

che tipi di query? in ordine di complessità:

- * query su metadati -> query esatta, è anche la più facile, si poteva rispondere a queste query anche con DBR classici: però ha i problemi del TAGGING, ovvero io posso taggare foto di bambino con automobile e poi questa mi viene restituita su query che cerca metadato AUTOMOBILE

* query con confronto di immagini -> "trovami tutte le immagini che assomigliano a questa", è più complessa perché comporta di entrare nel merito del concetto di SOMIGLIANZA, quali features devo considerare per la somiglianza? somiglianza del soggetto, dei colori?
in generale si vogliono ottenere i risultati migliori, perché tutte le immagini assomiglieranno con un certo grado anche minimo alla query: quindi mi interessano solo i risultati più interessanti, mi concentro solo su un sottoinsieme dei dati del db
* query by sketch -> anziché dare un'immagine dello stesso tipo dell'img che si cerca, faccio uno sketch: es per trovare un viso disegno ovale con occhi, naso, bocca
* query su sottoimmagini -> "immagini che contengono una parte che assomiglia ad img data"

=====

* query su contenuto -> "img che contiene una macchina"
* "img che contiene sia macchina sia uomo che assomiglia a quello nella foto segnaletica": più complicato!
* tutte le coppie di immagini che contengono oggetti simili, per conoscere regolarità di pattern

=====

abbiamo bisogno di linguaggio SQL-like!

query di esempio:

"select image P, object o1, object o2..."
trova immagine che contiene o1 e o2 e rispetta particolari vincoli (monte Fuji sopra al lago)

* oggetto1 è verosimilmente una montagna e visivamente assomiglia al contenuto dell'img data
* oggetto2 è verosimilmente un lago e visivamente assomiglia al contenuto dell'img data
* i due oggetti sono in una certa relazione spaziale

sono tutti aspetti non trattati dalle BDR!

=====

con questo linguaggio SQL-like posso esprimere la query!

rappresentazione visuale: montagna è contenuto di Fuji_mountain.gif, lago è contenuto di lake_image_sample.gif, e con freccia rappresento posizione degli oggetti nell'immagine

il sistema, a fronte di query complesse di questa natura, risolve individualmente i risultati della query e li combina per dare il risultato globale

=====

query and results:

- 1) il lago matcha e anche la montagna, però non vale il soddisfacimento della relazione spaziale
- 2) lago e montagna matchano parzialmente, relazione spaziale invece matcha perfettamente
- 3) manca il lago e il matching di montagna e relazione spaziale sono soddisfatti solo in parte (direzione può essere misurata in base al baricentro)
- 4) soddisfatta totalmente la componente spaziale, al 50% invece montagna e lago

QUALE MATCHA MEGLIO?

dipende! è risultato soggettivo che dipende dall'importanza data alle diverse componenti della query!

quale funzione uso per combinare i risultati? media, min, max?

col min, per esempio, perdo i match parziali, perché basta che una componente sia a 0 e l'oggetto non viene considerato!

RELEVANCE FEEDBACK: l'utente dà feedback sulla rilevanza delle immagini restituite come risultato, e il sistema apprende quali possono essere le ragioni per cui l'utente preferisce un'immagine piuttosto che un'altra: ad es. se l'utente sceglie 1), allora il sistema inferisce che l'utente dà più importanza alla presenza degli oggetti piuttosto che alla collocazione spaziale, quindi deve rifare la query pesando diversamente le componenti dando priorità diversa!

spesso l'utente stesso non ha bene idea inizialmente di quello che cerca quindi può voler raffinare la query

=====

DBM: partiamo da zero o usiamo modelli già esistenti?

consideriamo prima le BD relazionali e vediamo pro/contro:

- * rappresento dati con strutture elementare: stringhe di testo o numeri, quindi ci limitano perché non rappresentano immagini! su questo si basano strategia di memorizzazione, processazione query e ottimizzazione
- * dati rappresentati come tabelle, schema descrive il contenuto della tabella, chiavi di tabella, attributi con dominio

vogliamo mutuare caratteristiche principali a BDM:

- disco diviso in unità logiche
- relazione memorizzata in pagine contigue
- struttura ad indice permette di accedere ai record di una pagina in tempo logaritmico: ad ogni passo scende di un livello nell'albero -> riduce numero di accessi alla memoria
---> vogliamo qualcosa del genere anche per i DBM: vogliamo trovare organizzazione dei surrogati che ci permetta di arrivare ai dati che ci interessano senza esplorare l'intera base di dati!

=====

vogliamo query language dichiarativo:

diciamo cosa vogliamo ma non come ottenerlo -> permette ottimizzazione della query

vogliamo linguaggio dichiarativo perché se facciamo scegliere all'utente la strategia, lui presumibilmente non sceglie quella ottimale! il sistema deve esonerare l'utente da questo compito e demandare al modulo di ottimizzazione la pianificazione della strategia da adottare

BDM: sono statiche, non c'è molta dinamicità e i tempi di ricerca sono generalmente superiori

=====

STORAGE MANAGEMENT

i confronti vengono fatti sui surrogati, l'immagine viene restituita solo alla fine! c'è solo un puntatore all'immagine vera e propria.

prima vengono estratte le features tramite l'analisi del dato e memorizzate opportunamente

QUERY MANAGEMENT

traduce la query e specifica come devono essere visualizzati i risultati, in quale ordine: nell'ordine in cui sono state trovate oppure in un altro ordine specificato dall'utente (mentre nel DBR il risultato è una tabella)

QUERY PROCESSOR

ottimizzazione delle query in base al tradeoff costo-qualità: voglio qualità maggiore oppure costo minore? inoltre vogliamo risultati ordinati in base ad un rank, vogliamo vedere prima i risultati più rilevanti

POST QUERY PROCESSING

fase di raccolta feedback utente per capire se i risultati siano proprio quelli che l'utente si aspettava

=====

1) creazione del contenuto, nel senso di renderlo disponibile nella BD. comporta una serie di passi:

fase di upload del documento nella BD con

- registrazione del dato per renderlo disponibile, prevede verifica che il dato non esistesse già nella BD, e se c'è già, aggiornare?
- nel caso di aggiornamento si identificano le modifiche e queste vengono registrate eventualmente in questa fase viene effettuato watermarking: si inserisce una qualche sorta di filigrana o chiave che trasforma il documento e ha come obiettivo l'individuazione della proprietà del documento (per evitare violazione di copyright), all'occhio umano non dovrebbe accorgersi del fatto che l'immagine è stata trasformata per portare la firma del proprietario (unico detentore della chiave per riportare l'immagine al suo formato originale)

2)

- dato è oggetto di una fase di analisi per riconoscere ed estrarre le features, ovvero le caratteristiche che verranno memorizzate nel surrogato: vengono estratte caratteristiche sulla struttura, es: immagine composta di altre immagini
- ciascuna di queste feature verrà indicizzata da un'opportuna struttura: alcune strutture indice gestiscono l'intero documento, mentre altre gestiscono una cerca feature i cui risultati vengono poi combinati

3-4) query: lavoro di analisi sulla query che deve essere espressa nel modello in cui sono espressi i dati; nel fare questo possono servire

- info sull'utente per disambiguare
- query viene descritta in un qualche linguaggio
- query gestita da moduli che la rendono nella forma corretta per il modello in cui sono rappresentati i dati

5) formulazione query in pasto al modulo che gestisce le query di somiglianza: restituire oggetti più simili alla query. esistono due categorie di query:

- query k-results: primi k risultati più simili, quindi pone limite su cardinalità
- query di range: pongono limite su massima distanza dalla query, non su cardinalità

6) - aggancia i dati al surrogato del db: alla fine all'utente viene restituito il dato, quindi questa info deve essere accessibile: interviene solo nella restituzione del risultato (non nella fase di risoluzione della query)

7) feedback utente può modificare la query e quindi rieseguirla ottenendo nuovi risultati

=====

perché non vanno bene le basi di dati ER?

- DATI: solo testo o numeri -> non consente di memorizzare direttamente immagini, posso al massimo registrare il nome dell'immagine che viene memorizzata separatamente
- OPERATORI DI CONFRONTO: >, <, =, != ma non ho operatori per confronto di immagini e anche di MATCH PARZIALE che mi permette di associare grado di plausibilità al match e quindi di fare query NON ESATTE (nel modello ER oggetti o sono uguali o diversi, non c'è concetto di match parziale)
- MANCA RANKING: non ordina i risultati in modo che il primo sia quello che meglio soddisfa la query
- MANCA POTENZA COMPUTAZIONALE, manca analisi delle immagini

==> modello inadeguato

=====

soluzioni?

- approccio relazionale: usare linguaggio host e query embedded nel linguaggio
- approccio OO: consente più potenza computazionale

=====

modello ad oggetti è sufficiente?

- cattura gerarchie di aggregazione e di ereditarietà
- * AGGREGAZIONE: permette di definire un oggetto/concetto come aggregazione di altri oggetti/concetti -> figlio è parte del genitore
- * EREDITARIETÀ: riguarda informazione tassonomica -> figlio è anche genitore voglio rappresentare entrambe queste gerarchie nelle DBM! OODB mi danno questa possibilità

=====

OODB:

modellano meglio il mondo reale
permettono di definire propri metodi

PROBLEMI:

- adatto ad applicazioni di business
- un po' adatto a gestione di basi multimediali (nuovo Oracle gestisce un po' somiglianza di oggetti multimediali)
- query esatte
- query procedurali
- privilegiati aspetti legati a concorrenza/updates, legata ancora ad aspetti dinamici ma grossi archivi di dati multimediali sono più statici

SVANTAGGI:

- difficile ottimizzazione
- no match parziale
- manca ranking
- query processing è guidata dal costo di valutazione, non dalla somiglianza

=====

ORDB:

prendono aspetti di entrambi:
da R -> tuple e linguaggio SQL
da OO -> funzioni definite dall'utente e ADT
ma anche questo non va bene: i punti di debolezza comuni ci sono ancora...

=====

ci sono anche altri modelli di DB:

- * BD deduttive: la conoscenza è espressa sotto forma di clausole logiche (tipo quelle di Prolog)

CARATTERISTICHE

- + basati su logica, ci servono questi aspetti per combinare risultati di matching parziali (congiunzione di risultati di sotto-proprietà)
- query booleane

* DB fuzzy: grado di soddisfacimento delle query è sfumato (non binario), ogni risultato è associato ad un grado di soddisfabilità: ovvio perché c'è troppa imperfezione nel surrogato

CARATTERISTICHE

- + logica non booleana
 - + risultati non esatti -> matching parziale
- * BD spazio-temporali: pensate per gestire concetti spaziali o temporali, gestire evoluzione dati nel tempo

CARATTERISTICHE

- + query di range o nn
- + modello basato su vettori o intervalli
- + query dichiarative

* data mining: nel momento in cui associamo semantica ad un'immagine dobbiamo trovare pattern che si ripetono, apprendere regole di comportamento dal pattern, classificare oggetti

* sistemi di gestione di dati semi-strutturati: XML

- oggetti si auto-descrivono

* BD di immagini

Vectors...what are they???

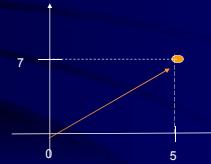
- Image with 1 pixel <5>



Maria Luisa Sapino - Basi di dati
Multimediali

Vectors...what are they???

- Image with 2 pixels <5,7>



Maria Luisa Sapino - Basi di dati
Multimediali

Vectors...what are they???

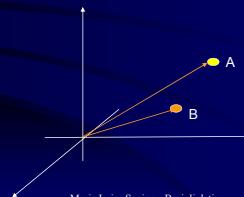
- Image with 3 pixels <5,7,3>



Maria Luisa Sapino - Basi di dati
Multimediali

Distance between two images???

- Given $A< a_1, a_2, a_3 >$ and $B< b_1, b_2, b_3 >$, how different are they?



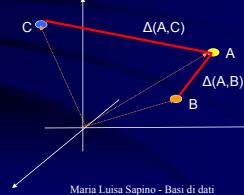
Maria Luisa Sapino - Basi di dati
Multimediali

Euclidean distance

$$\Delta(A, B) = \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2 + (a_3 - b_3)^2}$$

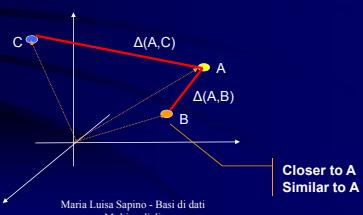
Maria Luisa Sapino - Basi di dati
Multimediali

Which image is more similar to A?

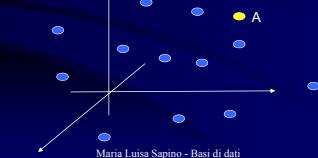


Maria Luisa Sapino - Basi di dati
Multimediali

Which image is more similar to A?



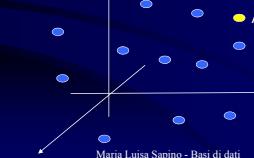
“Find 2 most similar images to A”



“Find 2 most similar images to A”

Maria Luisa Sapino - B:
Multimediali

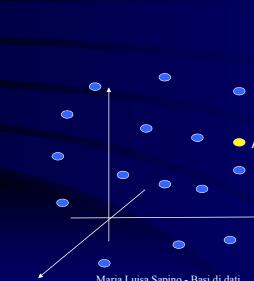
“Find images at most δ different from A”

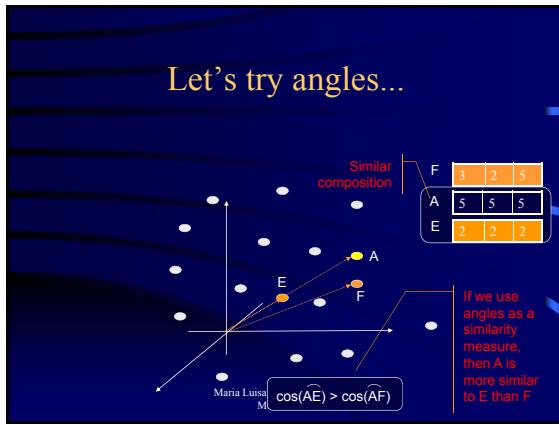
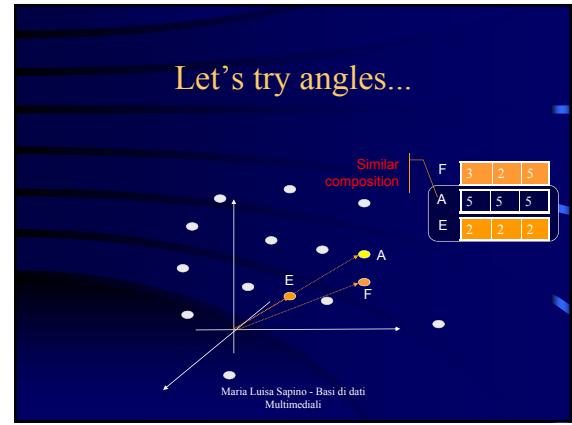
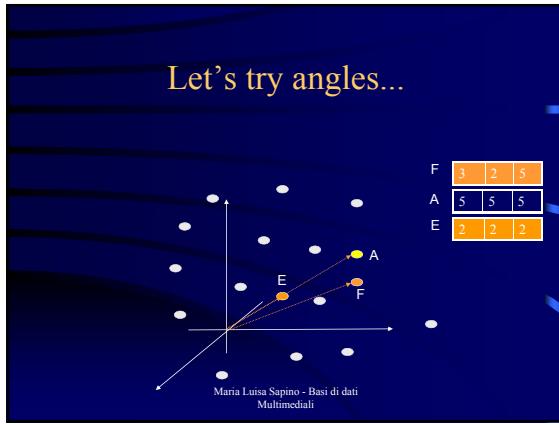


Find images at most δ different from A

Maria Luisa Sapino - B:
Multimediali range search

Are there other similarity measures?





Angle-based measures

- Given $\vec{x} = \langle x_1, x_2, \dots, x_n \rangle$ $\vec{y} = \langle y_1, y_2, \dots, y_n \rangle$
- Dot product $\vec{x} \cdot \vec{y} = \sum_{i=1}^n x_i y_i$
- Cosine similarity $\cos(\vec{x}, \vec{y}) = \frac{\vec{x} \cdot \vec{y}}{\|\vec{x}\| \|\vec{y}\|}$

Maria Luisa Sapino - Basi di dati Multimediali

What is a good measure then??

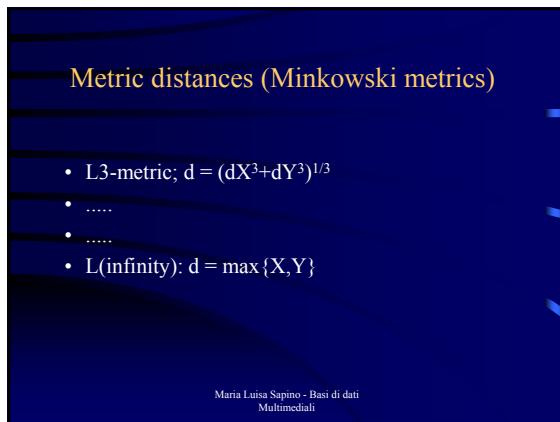
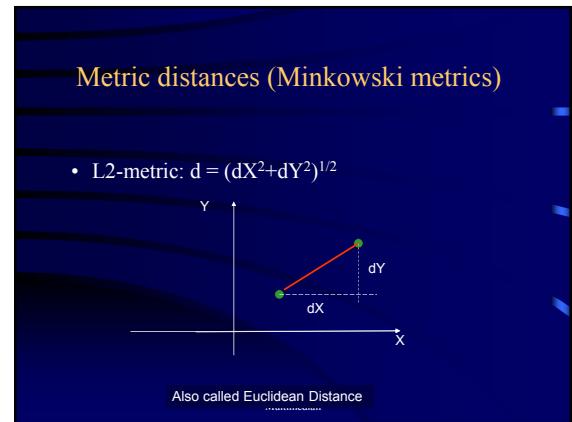
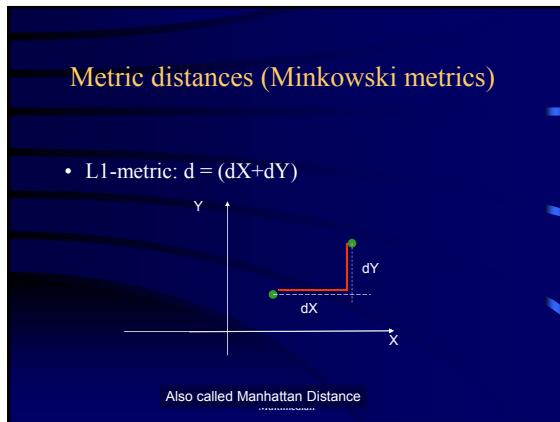
- Application dependent...
- ...but, distances in a metric space help indexing!

Maria Luisa Sapino - Basi di dati Multimediali

Metric model: axioms

- Any function d expressing a distance must satisfy the following axioms:
 - self-minimality: $d(s, s) = 0$
 - minimality $d(s_1, s_2) \geq d(s_1, s_1)$
 - symmetry $d(s_1, s_2) = d(s_2, s_1)$
 - triangular inequality $d(s_1, s_2) + d(s_2, s_3) \geq d(s_1, s_3)$
- Example: Euclidean distance

Maria Luisa Sapino - Basi di dati Multimediali



...metric model

- Well suited for certain kinds of similarity evaluation, such as color based comparisons
- Consistent with widely used approaches from computer vision and pattern recognition communities
 - results suggest that the L1 metric may better capture human notions of image similarity.
- Makes it relatively easy to index data, modeled as vectors of properties, in terms of classical multi-dimensional indexing techniques.

Maria Luisa Sapino - Basi di dati
Multimediali

DEFINIZIONE DEL MODELLO DEI DATI:

quale modello utilizzare per rappresentare i dati? quale tipo di struttura per contenere l'informazione?

uno dei più utilizzati è il MODELLO VETTORIALE: per rappresentare dati di tipo multidimensionali, oggetti le cui caratteristiche sono definite da più dimensioni dove ciascuna dimensione corrisponde ad un valore che una feature rappresentata da una dimensione può assumere: per le immagini ad es. si può associare un elemento a ciascun pixel dell'immagine e dargli come valore il colore del pixel -> confrontare immagine = confrontare vettore dell'immagine

se abbiamo due oggetti A e B rappresentati in uno spazio tridimensionale, il confronto viene fatto sul valore dei vettori: come definisco questo confronto?

1) DISTANZA EUCLIDEA: radice della somma della differenza tra ciascuna coppia di coordinate al quadrato

visivamente B è più simile ad A (rispetto alla somiglianza euclidea) -> immagine B è più simile ad A

=====

come rispondere a query NN?

trovare gli N punti più vicini: devo avere indice che permetta di organizzare i dati nello spazio per permettermi di potare alcune parti dello spazio

come rispondere a query di range?

trovare i punti che distano al più \delta dalla query: devo costruire (se uso distanza euclidea) un'ipersfera attorno alla query e restituire tutti i punti che cadono all'interno di questo spazio: di nuovo l'indice deve consentirmi di potare il più possibile

=====

per certe applicazioni (es: TESTO) non va bene distanza euclidea!

(let's try angles - pag.3 lucido 1)

ANGOLI: quale caratteristica considero per calcolare la distanza? in questo caso
- per distanza euclidea A è più simile a F
- ma A ed E hanno componenti che hanno le stesse reciproche proporzioni

ANGOLI catturano più somiglianza di composizione interna, quindi meglio per i testi perché controllo il contenuto -> non misuro la lunghezza del vettore differenza, ma l'ampiezza dell'angolo tra i due vettori

* MISURA COSENO: quantifica la somiglianza in termini di misura del coseno dell'angolo compreso tra i due vettori

=====

quale misura? nessuna è meglio dell'altra in assoluto: dipende dall'uso che devo farne!

- se devo misurare se un oggetto è un "riassunto dell'altro" uso coseno (non euclidea perché altri mettono anche la dimensione)

dot product: più alto se i punti sono più vicini

somiglianza coseno: normalizza rispetto alle lunghezze dei vettori

N.B. -> distanza coseno = 1 - somiglianza coseno

=====

la scelta della misura ha impatto sul tipo di indice da usare: perché l'indice per fare delle potature deve decidere di scartare parti di spazio da esplorare o no in base alla misura della distanza

=====

indipendentemente dall'applicazione, per poter definire indici basati su queste misure che permettano poi di potare, la misura deve essere una METRICA, ossia deve soddisfare alcuni requisiti:

- 1) self-minimality: quando io misuro la distanza tra un oggetto e sé stesso questa deve essere 0 -> ogni oggetto è uguale sé stesso
- 2) minimality: distanze sono solo positive o nulle
- 3) simmetria: distanza tra s1 e s2 = distanza tra s2 e s1
- 4) disegualanza triangolare: somma dei due lati di un triangolo è \geq del terzo -> distanza tra due oggetti non può superare la somma tra la distanza tra ciascuno dei due oggetti e un terzo

=====

DISTANZA DI MINKOWSKI:

L1 -> distanza di Manhattan cattura che la distanza tra due punti equivale alla somma delle differenze tra le singole coordinate -> è il più vicino al modo di pensare dell'utente (ranking utente intuitivamente fatto in base a manhattan)

L2 -> distanza euclidea

L3 ecc ecc

L(\infty) -> massimo tra i due oggetti

=====

es. di distanza non metrica: edit distance, cioè quanto costa trasformare una stringa in un'altra con operazioni elementari (cancella, inserisci, sostituisci) -> se dò costi diversi alle operazioni, es. inserire costa di più che cancellare, allora non vale più la simmetria:
es. distanza tra NONNO e NONO è diversa della distanza tra NONO e NONNO

quelle che usiamo per costruire indici sono tutte metriche

Feature...

- ...a property of interest that can help us index an object
- For a “student record”
 - student_IDcan be a feature
- What are the features for an image?

13
8

Maria Luisa Sapino (BDMM 2011)

Image features

- There are many possible features
 - Color histogram
 - Texture
 - Edges
 - Shapes
 - Objects
 - Object or scene semantics
- Feature selection: which one to use for indexing?

13
9

Maria Luisa Sapino (BDMM 2011)

Good feature..

- A good feature is **significant** and enables us to **differentiate** objects from others as much as possible
- A good feature corresponds to users' perception as much as possible
 - Relevance feedback!!!!

14
0

Maria Luisa Sapino (BDMM 2011)

What does “significant” mean

- Information theoretic sense:
 - An event is more significant if it carries more information

14
1

Maria Luisa Sapino (BDMM 2011)

What does “significant” mean

- Information theoretic sense:
 - An event is more significant if it carries more information
 - An event that has high occurrence rate carries less information
 - Solar eclipse is more interesting than sunset

High frequency ---- less information
Low frequency ---- high information

14
2

Maria Luisa Sapino (BDMM 2011)

Entropy

- Total information content (uncertainty)

$$H(X) \equiv \sum_{x \in \mathcal{A}_X} P(x) \log \frac{1}{P(x)}$$

14
3

Maria Luisa Sapino (BDMM 2011)

Entropy (example)

- Total information content (uncertainty)

$$H(X) \equiv \sum_{x \in \mathcal{A}_X} P(x) \log \frac{1}{P(x)}$$

$P(a) = 0.5, P(b) = 0.5 \rightarrow H > 0$ more uncertain

$P(a) = 1.0, P(b) = 0.0 \rightarrow H = 0$ less uncertain

14
5

Maria Luisa Sapino (BDMM 2011)

Entropy (example)

- Total information content (uncertainty)

$$H(X) \equiv \sum_{x \in \mathcal{A}_X} P(x) \log \frac{1}{P(x)}$$

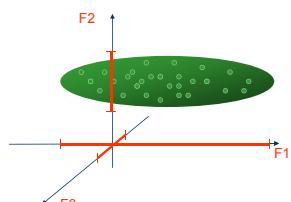
$P(a) = 0.5, P(b) = 0.5 \rightarrow H > 0$ more uncertain
more information

$P(a) = 1.0, P(b) = 0.0 \rightarrow H = 0$ less uncertain
less information

14
6

Maria Luisa Sapino (BDMM 2011)

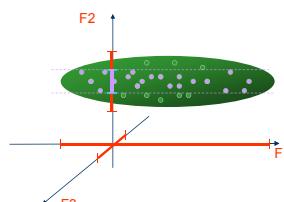
Which feature is better?



14
7

Maria Luisa Sapino (BDMM 2011)

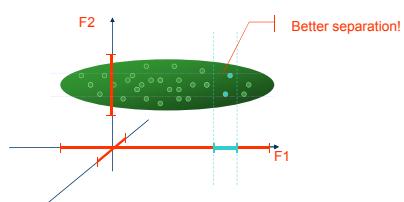
Which feature is better?



14
8

Maria Luisa Sapino (BDMM 2011)

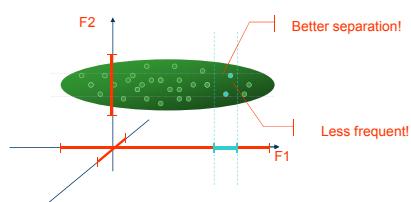
Which feature is better?



14
9

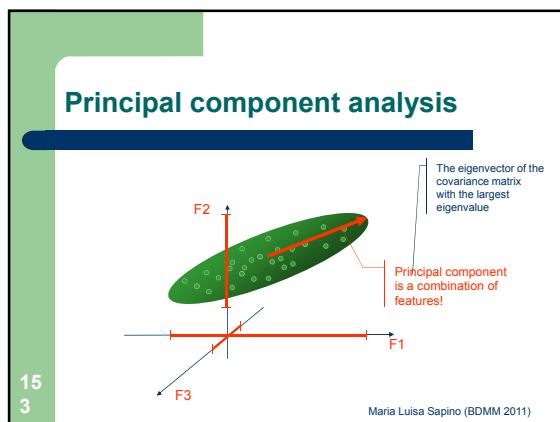
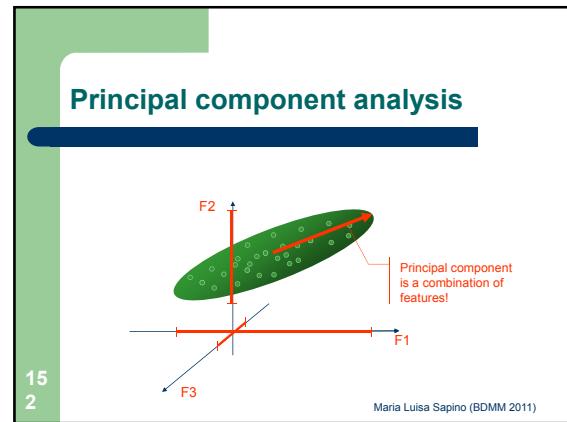
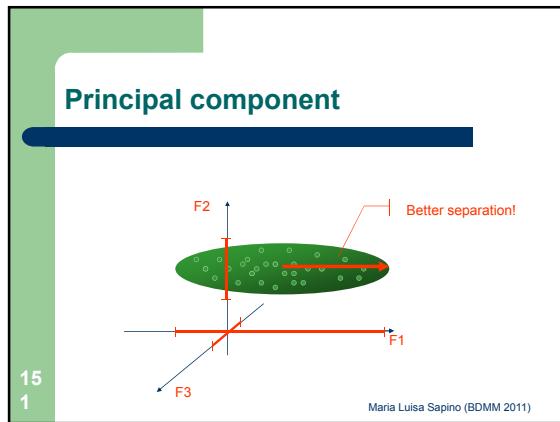
Maria Luisa Sapino (BDMM 2011)

Which feature is better?



15
0

Maria Luisa Sapino (BDMM 2011)

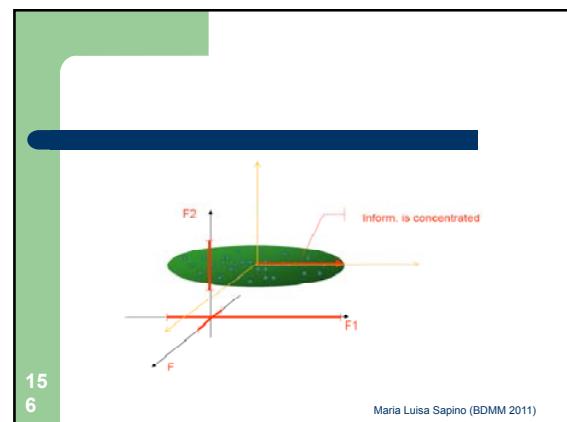
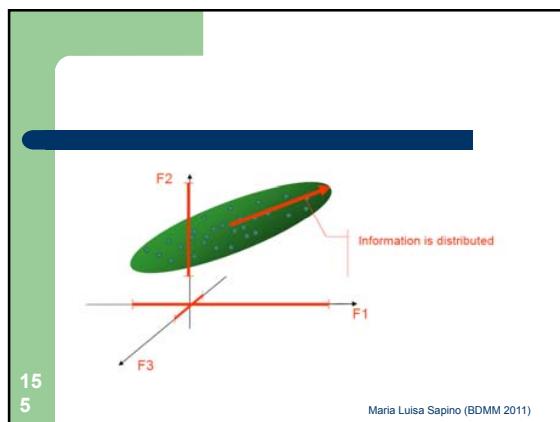


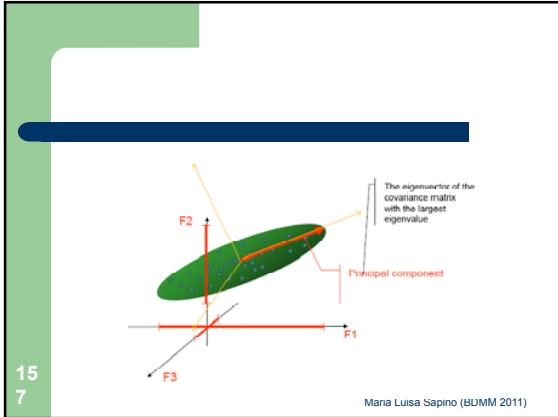
Principle Component Analysis

- ..also known as Karhunen-Loeve Transform
 - ..a linear transform that optimally decorrelates the input.

15
4

Maria Luisa Sapino (BDMM 2011)



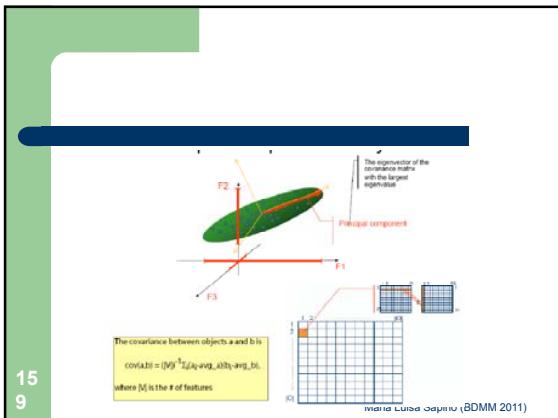


Linearly Independent Eigenvectors

- Suppose that A is an $n \times n$ square matrix. If the eigenvalues, $c_1 \dots c_k$ are distinct, then eigenvectors v_1, \dots, v_k are a set of k linearly independent vectors.

15
8

Maria Luisa Sapino (BDMM 2011)

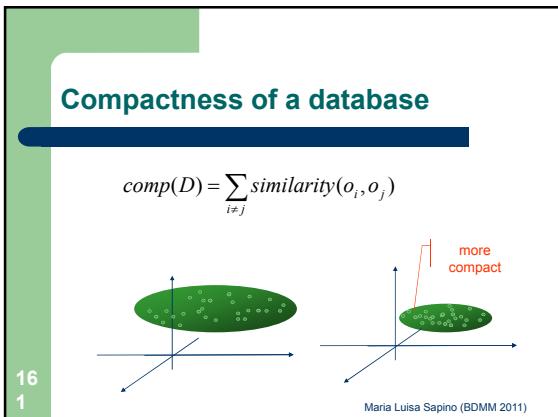


Compactness of a database

$$\text{comp}(D) = \sum_{i \neq j} \text{similarity}(o_i, o_j)$$

16
0

Maria Luisa Sapino (BDMM 2011)



Compactness of a database

$$\text{comp}(D) = \sum_{i \neq j} \text{similarity}(o_i, o_j)$$

A compact database is not desirable!!!

16
2

Maria Luisa Sapino (BDMM 2011)

Feature quality

A feature is
 - good if we remove it, the overall compactness increases
 - bad if we remove it, the overall compactness decreases

16
3

Maria Luisa Sapino (BDMM 2011)

Problem...

- Feature vector size: 628×1024
 - Dimensionality curse: high dimensions make indices unusable (10-15 dimensions max!!!)

16
4

1024
628

Maria Luisa Sapino (BDMM 2011)

Problem...

- Feature vector size: 628×1024
 - Dimensionality curse: high dimensions make indices unusable (10-15 dimensions max!!!)
- Solution: Reduce # dimensions of the vector
 - use distance-preserving transforms
 - Ex: fourier trans., DCT, wavelet trans.

16
5

628 x 1024 DCT 4

Transforms

16
6

Maria Luisa Sapino (BDMM 2011)

Transforms

16
7

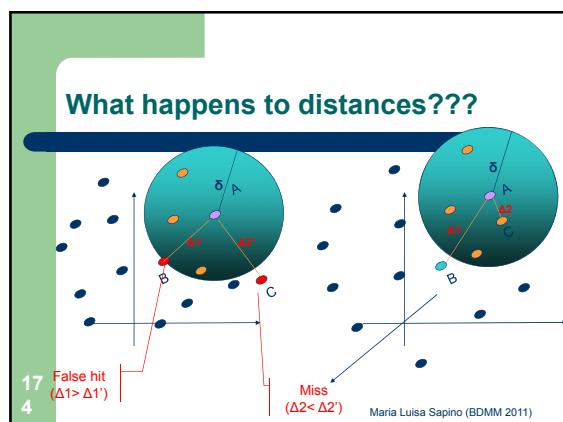
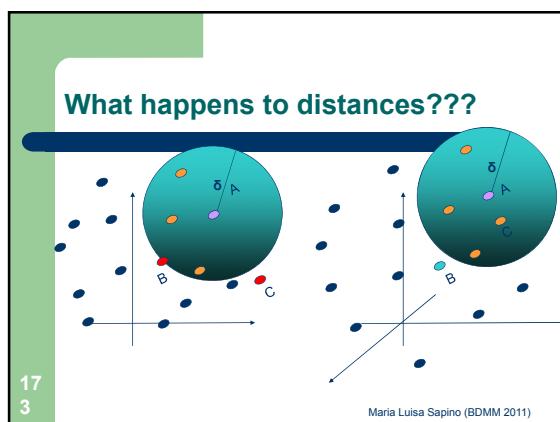
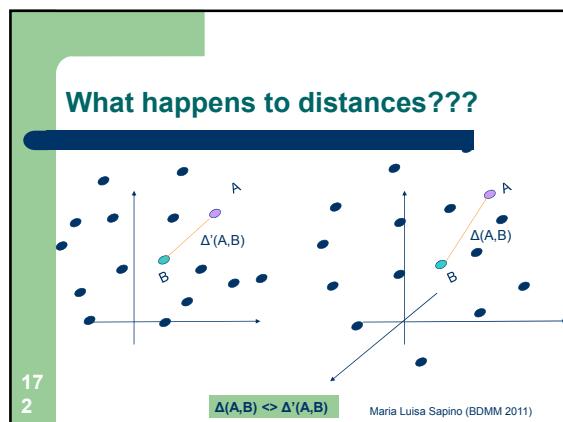
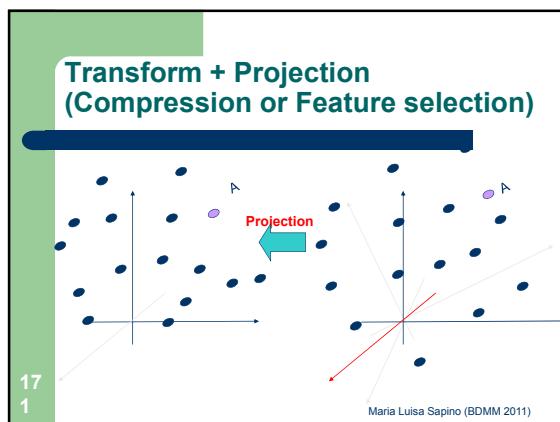
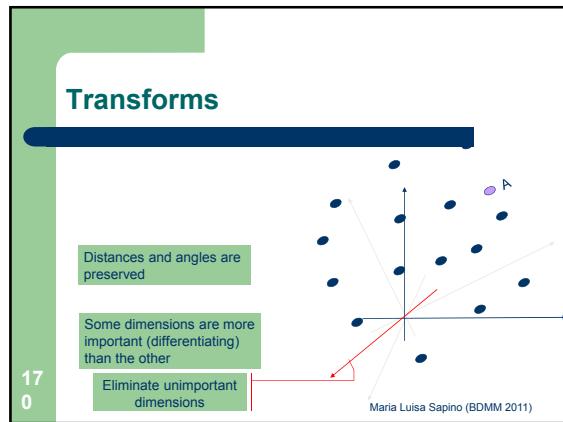
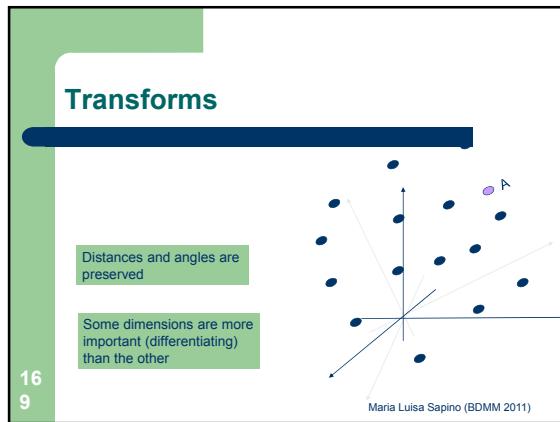
Maria Luisa Sapino (BDMM 2011)

Transforms

Distances and angles are preserved

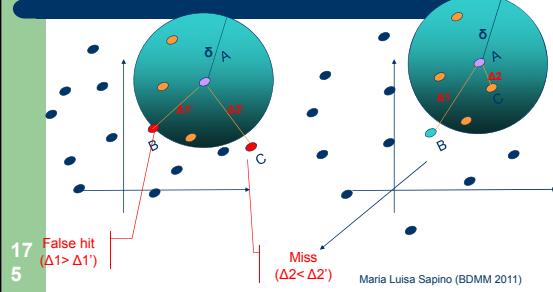
16
8

Maria Luisa Sapino (BDMM 2011)



Misses are not desirable!
Can not be eliminated with postprocessing

What happens to distances???



Maria Luisa Sapino (BDMM 2011)

dati multimediali: dati complessi per cui non riusciamo a rappresentare tutto, sia perché troppo complessi sia perché contengono aspetti soggettivi -> l'oggetto non può essere direttamente quello su cui andrò a fare confronti -> devo ricorrere ad un surrogato, cioè il modello che uso, l'astrazione che io faccio sul dato e su cui mi baso per fare le mie elaborazioni

N.B. poi all'utente restituisco il dato vero!
ma per scegliere quale dato restituire io lavoro sul surrogato

abbiamo visto le misure per le distanze (euclidea, coseno): rappresentare nel modello vettoriale in cui a ciascuna dimensione corrisponde una feature o un range di valori di una feature

possiamo memorizzare istogramma, texture, metadati di un'immagine

MA COME SCEGLIAMO LE FEATURES? su quale criteri ci basiamo per selezionarle?

=====

COS'è una FEATURE?
proprietà interessante nel caratterizzare l'oggetto e che può essere utile osservare ai fini dell'indicizzazione dell'oggetto
se scarto una feature non potrò più fare query su quella proprietà: è molto vincolante! è fondamentale nella progettazione del DBM analizzare bene le features per capire quelle che ci serviranno per fare le query

=====

PER UN'IMMAGINE (da più basso a più alto livello):

- istogramma dei colori
- texture (sensazione di rugosità, filigrana, regolarità del pattern)
- edges = contorni, separano aree con colori diversi
- forme (quadrato, cerchio, ecc...)
- oggetto nell'immagine (cane, persona)
- semantica dell'oggetto e delle scene (es. bambino che gioca a pallone)

QUALI USARE PER INDICIZZARE?

scegliere significa poter giudicare il contributo che, se presenti, queste features possono portare al DBM: se non le userò mai è meglio non includerle

=====

una feature BUONA:

- * deve essere significativa, ossia portare informazione, ed essere il più possibile differenziante, ossia che mi permetta di discriminare il più possibile due oggetti diversi (altrimenti la risposta ad una query potrebbe restituire tutto il contenuto del db!)
- * deve essere il più possibile vicina a quello che l'utente percepisce dell'oggetto, altrimenti non la chiederà mai e quindi è inutile che io la memorizzi: importante anche per il relevance feedback, perché l'utente si basa su features che percepisce per valutare

=====

FEATURE SIGNIFICATIVA: porta informazione, ovvero è relativamente rara ed imprevedibile -> se è meno comune allora è più significativa

le features i cui valori sono meno scontati sono più significative, e quindi migliori candidate ad essere memorizzate

=====

COME CALCOLARE l'INFORMAZIONE?

calcolo dell'entropia che misura il grado di incertezza del sistema: entropia alta -> molta informazione, entropia bassa -> poca informazione

- * se ho due eventi, uno certo e uno che non accadrà mai, ho la minima entropia
- * al crescere dell'incertezza cresce l'entropia
- * se è meno prevedibile porta più informazione

ragione per scegliere features: DIMENSIONALITY CURSE, per cui se io ho un surrogato e indicizzo

rispetto ad un surrogato che ha più di 10-15 features, il costo di una ricerca su indici supera il costo della ricerca sequenziale -> una query deve consentire di restituire un insieme limitato di oggetti, non l'intero DB!

=====

ES (figura pag. 2)

modello vettoriale: ho salvato i miei surrogati memorizzando 3 features, es. componenti RGB

questi dati sono distribuiti in questo spazio che è "allungato" rispetto all'asse orizzontale: se io dovesse scegliere se lasciar cadere F1 o F2, mi converrebbe tenere F1 perché è più differenziante.

come lo misuro? prendo una certa unità di misura su ciascuna delle due features e vedo quanto è discriminante: segmento su F2 include quasi tutti gli oggetti, mentre lo stesso segmento su F1 include molti meno elementi (meno frequenti) -> F1 dà una maggior separazione dei valori e quindi è preferibile rispetto ad F2, perché nella stessa interrogazione su stessa distanza con features diverse, F2 mi restituirebbe molti più risultati!
ricadere nello stesso intervallo su F1 è molto meno probabile che ricadere nello stesso intervallo su F2

è bene che gli oggetti siano separati: questo ha impatto sull'efficacia dell'indice: se l'indice è tale per cui a fronte della stessa query non si poteva niente (restituisce tutta la BD), allora il costo del mantenimento dell'indice non è compensato dal beneficio che dovrebbe derivarne

=====

cosa succede se la componente di massima separazione non è allineata con una delle tre features?

nell'es: la max separazione è obliqua

in questo caso si deve fare una TRASFORMAZIONE DI SPAZIO, ossia ci si riconduce ad una rotazione degli assi che porta il sistema di riferimento ad essere allineato alla direzione di massima separazione (obliqua): questo ricondursi a questo spazio nuovo si chiama PRINCIPAL COMPONENT ANALYSIS

la componente di massima separazione è data dall'autovettore della matrice di covarianza avente il più alto autovalore:

* oggetto rappresentato in n dimensioni con [o_1, o_2, ..., o_n]

* v_i = media di tutti i valori degli oggetti per la dimensione i

* date due dimensioni i e j, la COVARIANZA mi quantifica il modo in cui queste due dimensioni variano insieme, sono correlate:

$\sigma_{i,j} = 1/n * \sum_{h=1}^n (x_{hi} - v_i)(x_{hj} - v_j)$

* matrice delle covarianze: matrice quadrata che per ogni cella $i..n$ $j..n$ mi dice la covarianza di $\sigma_{i,j}$, se tutte le dimensioni fossero completamente indipendenti la cov sarebbe 0 (tranne sulla diagonale), in ogni caso è una matrice simmetrica

* data matrice quadrata C, è sempre possibile decomporla in un prodotto di 3 matrici B*A*C: la A è una matrice (quadrata) diagonale con $r \leq n$, ha tutti 0 tranne sulla diagonale i cui valori sono ordinati in modo decrescente, e rappresentano gli AUTOVALORI: l'autovalore più alto dice quale è la componente principale rispetto alla quale sono distribuiti i valori, ma anche la MENO importante ovvero quale dovrei fare cadere

=====

altro parametro per giudicare: COMPATTEZZA

misura la somiglianza globale -> per ogni coppia di oggetti distinti del DB, misura la somiglianza tra i due e somma tutte le somiglianze reciproche -> un DB è tanto più compatto quanto gli oggetti sono simili, corrisponde visivamente a punti molto vicini nello spazio: voglio massimizzare la separazione e MINIMIZZARE la compattezza -> se il DB è molto compatto, se io faccio query di range vengono restituiti troppo oggetti -> poco potere discriminante

=====

+) features buone: se rimosse, fanno aumentare la compattezza del DB -> servono a mantenere il DB poco compatto

-) features cattive: se rimosse, il DB rimane poco compatto

=====

dimensionality curse:

suddivisione gerarchica dello spazio che permette di scendere sul cammino e potare altri rami essendo certi che su quel ramo non c'è risultato cercato -> con dimensionality curse non c'è questa certezza perché i figli hanno regione di overlapping, quindi devo seguire tutti i percorsi -> al crescere delle dimensioni aumenta l'overlapping e devo percorrere tutti i rami, quindi l'accesso ai dati non si riduce alla visita di un cammino ma comporta esplorazione completa dell'albero -> molto costosa!!

COME RIDURRE LO SPAZIO?

- * scegliamo le features con cura
- * ma è possibile che le features siano tutte abbastanza simili da quel punto di vista e quindi non abbia elementi per sceglierne una piuttosto che un'altra

allora COME FARE?

=====

primo modo: partire da un vettore con molte dimensioni e ridurre le dimensioni del vettore, es. usando trasformazioni che permettono di ottenere stesso risultato che otterrei su più dimensioni

* ruotare spazio di partenza in un certo modo: le distanze reciproche e gli angoli non cambiano
->
dopo la rotazione posso riconoscere che a questo punto certe dimensioni sono più importanti di altre! quindi posso eliminare le dimensioni meno significative

* se invece faccio anche una proiezione, ovvero faccio cadere una dimensione, faccio una trasformazione che non è più senza perdita: scelgo la dimensione che meno differenzia gli oggetti e la lascio cadere -> gli oggetti si avvicinano quanto meno è differenziante la dimensione -> eliminando una dimensione faccio una proiezione sulle altre (perché la prima dimensione va a 0)

N.B. oggetti possono anche allontanarsi!

=====

nel mio spazio trasformato ho due oggetti A e B che distano \Delta: cosa succede quando proietto i due oggetti?
* la distanza cambia in '\Delta' che è plausibilmente diversa (a meno che A e B non avessero lo stesso valore lungo quella direzione)
* se faccio una query di range, idealmente vorrei che tutti gli oggetti che stavano al più \delta fossero compresi (voglio restituire gli stessi oggetti di prima! le query non devono risentire della trasformazione che ho fatto); in realtà però è possibile che ora B non venga più compreso nel range (MISS) o che qualche oggetto che prima era fuori ora sia compreso (FALSE HIT): in termini di danno un MISS è più grave -> infatti per ritrovare un oggetto "missed" dovrei rifare la query! invece un FALSE HIT può essere riconosciuto in una fase di post-processing (analisi completa su tutti i risultati) senza dover ripetere la query
=> post-processing invece non può recuperare oggetti persi!

* query di range: distanza tra oggetti nel loro complesso, non tra singole features -> la distanza si calcola sulle features indicizzate -> se ho indice multidimensionale allora calcolo distanza tra tutte le features indicizzate
* k-nn: pongo un limite al numero di oggetti che voglio come risultato -> raramente il ranking corrisponde esattamente a quello percepito dall'utente

====

SEMANTIC GAP: es. smilie a livello di estrazione di features di basso livello non sono interpretabili facilmente a livello semantico -> differenza tra significato che io associo a questa immagine e caratteristica visuale dell'immagine

RIASSUNTO DELLE PUNTATE PRECEDENTI:

- * complessità del dato multimediale che non può essere rappresentato in modo fedele
- * informazioni che estraiamo dagli oggetti multimediali sono FEATURES e concorrono alla definizione del surrogato dell'oggetto, ovvero una sua rappresentazione (o modello) che è conforme ad un certo standard che adottiamo (es: modello vettoriale)

Media and Features

MEDIA AND FEATURES

2 tipi di media che consideriamo:

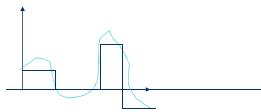
- 1) testo -> aspetti simbolici
- 2) immagini -> aspetti visuali

perché parliamo solo di immagini e testo?

sono due media sufficientemente diversi tra di loro,
quindi ci permettono di introdurre la maggior parte dei
concetti importanti in tema di multimedia:

Signal...

- Is a function (generally of time) $f(t)$



Used in representing analog and digital information
-analog signal → continuous
-digital signal → discrete

17
7

Audio: f : time → volume (analog)

Image: f : coordinate × coordinate → color. (digital)

Maria Luisa Sapino (BDMM 2010)

- Black and white images

– f : coordinate × coordinate → $\{0, 1\}$

- Greyscale images

– f : coordinate × coordinate → $[0, 255]$

- Color (Depth = 24)

– f : coordinate × coordinate → $[0, 255] \times [0, 255] \times [0, 255]$

- Color image (with color table)

– f : coordinate × coordinate → color index $[0, 255]$

– C_t : color index → $[0, 255] \times [0, 255] \times [0, 255]$

17
8

Maria Luisa Sapino (BDMM 2010)

Filter....

-is a function that transforms an input signal $f(v)$ into an output signal $g(v)$
 - Filter: $f(v) \rightarrow g(v)$
- Linear filter:
 - $f(av) \rightarrow a^*g(v)$
 - $f_1(v) + f_2(v) \rightarrow g_1(v) + g_2(v)$
- Space invariant filter
 - $\text{filter}(f(x+a)) = \text{filter}(f(x))$

$$\begin{array}{c} A + B = C \\ \downarrow \quad \downarrow \quad \downarrow \\ A' + B' = C' \end{array}$$

Maria Luisa Sapino (BDMM 2010)

17
9

Text vs. images

- Text
 - Symbolic
 - Artificial
 - Single meaning (reader independent ?)
 - Small storage requirements

- Images
 - Visual
 - Natural, artificial
 - Multiple meanings (viewer dependent)
 - Large storage requirements

Maria Luisa Sapino (BDMM 2010)

18
0

- TESTO:**
- di natura simbolica -> rappresento concetto con sequenza di caratteri
 - natura artificiale
 - meno ambiguo dell'immagine -> non sempre il significato è unico (es: poesia, suscita emozioni diverse anche se il testo è lo stesso) ma c'è cmq meno soggettività rispetto all'immagine
 - richiede meno spazio per essere memorizzato
- IMMAGINE**
- rappresentazione del concetto è visuale
 - sia naturali che artificiali
 - più significati -> più soggettivo -> più utile il relevance feedback
 - richiede più spazio

Example: Images...

- Convenient ways to store visual information
 - Bitmap:
 - 2D array of pixels.
 - each pixel contains color+illumination information
 - They have to be
 - processed and
 - analyzedto extract the information content

Maria Luisa Sapino (BDMM 2010)

18
1

Image operations

- Acquire
- Store
- Browse
- Query/QBE
- Process/Analyse
- Index
- Retrieve/request
- Display
- Compress
- Watermark
- Transmit
- Enhancement

18
2

Maria Luisa Sapino (BDMM 2010)

delle immagini devo rappresentare le features -> devo estrarre le features!

* devo introdurre una fase di preprocessing cioè di estrazione delle features che costituiscono il surrogato -> ovviamente non può essere fatta ogni volta, on the fly, ma va fatta una volta sola quando si inserisce il surrogato!

ci sono due tecniche/attività complementari che concorrono alla definizione del surrogato:

- 1) IMAGE PROCESSING: insieme di processi/tecniche che, data un'immagine, trasformano l'img di partenza in un'altra immagine che è fortemente correlata a quella di partenza per enfatizzare alcuni aspetti dell'immagine che renderanno più facile la successiva fase di analisi, es: rendere più nitido il contrasto tra due regioni adiacenti, così saranno più efficaci algoritmi che riconoscono forme; inoltre smorzano aspetti di rumore che non portano semantica
- 2) IMAGE ANALYSIS: parte da immagine eventualmente pre-processata e restituisce le features, es: descrizioni di forme nell'immagine, istogramma dei colori, descrizione della semantica riconosciuta nell'immagine (=> riconoscimento di pattern, es: riconoscere volto da occhi+naso+bocca)

Processing vs. analysis..

• Image processing



• Image analyzing



18
3

Maria Luisa Sapino (BDMM 2010)
Cat, man, ombrello...

Image processing

- Operators
 - Sharpening
 - Blurring
 - Rotating
 - Translating
 - Brightening
 - Cut/paste/resize
 - Warping

18
4

Maria Luisa Sapino (BDMM 2010)

OPERAZIONI DI IMAGE PROCESSING:

- * SHARPENING: rende più nitidi i contorni, mette a fuoco
- * BLURRING: contrario -> sfoca
- * ROTATING: ruota
- * TRANSLATING: trasla sugli assi
- * BRIGHTENING: illumina
- * CUT / PASTE / RESIZE: varie modifiche di taglio - ridimensionamento
- * WARPING: deforma

la più interessante è SHARPENING perché se un'img è a fuoco ed è nitida è più facile riconoscere i contorni e quindi permette di applicare poi tecniche di riconoscimento delle forme ed estrarre le feature corrispondente

FASE DI ANALISI: restituisce le componenti (candidate) del surrogato

- * **ISTOGRAMMA DI COLORI:** numero di pixel per ogni colore presente nell'immagine (contatore associato a ogni colore nell'immagine che viene incrementato per ogni pixel)
- * **TEXTURE:** tessitura, dà sensazione di granularità, di superficie
- * **EDGES:** contorni, per capire che c'è un bordo non mi basta il valore assoluto di un pixel ma serve sapere il colore dei suoi vicini
- * **SHAPES:** forme
- * **OBJECTS:** a livello più alto di astrazione -> riconosce pattern
- * **SEMANTICS:** significato
- * **DEPTH:** immagini tridimensionali

- alcune di queste features possono essere estratte pixel per pixel:
prendo un pixel, lo analizzo una volta sola, e memorizzo informazione

associata -> es: istogramma dei colori

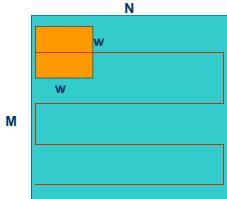
- invece tessiture, bordi, forme, posso capire se un pixel sta su un bordo solo se so che alla sua sx c'è un colore diverso rispetto a quello che c'è alla sua dx: il valore assoluto del pixel non mi dà informazione in questo senso! mi servono anche i valori dei pixel vicini! -> in questi casi è necessaria un'operazione di convoluzione di una matrice sull'immagine

Feature extraction

- Image analysis gives the features
 - Color histogram
 - Texture
 - Edges
 - Shapes
 - Objects
 - Semantics
 - Depth (stereoscopic images)
- Many of these operators require a filter to be convoluted over the original image

Maria Luisa Sapino (BDMM 2010)

Processing requirements...



Cost of the operation: $M \times N \times w \times w$
Maria Luisa Sapino (BDMM 2010)

faccio percorrere la matrice dell'immagine da una finestrella più piccola che si sposta limitando l'analisi dell'immagine ai pixel contenuti nella finestra

per EDGES: devo percorrere la matrice limitando l'analisi ai pixel contenuti nel quadrato arancione (CONVOLUZIONE)
non fa scansione elemento per elemento come per istogramma dei colori: qui considero la superficie quadrata e controllo se esistono discrepanze tra pixel adiacenti -> è operazione molto più costosa della precedente

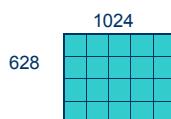
istogramma: costo $M \times N$ se matrice ha dimensione $M \times N$ perché controllo ogni pixel una volta sola

(N.B. ovviamente la finestra quadrata è molto più piccola rispetto alle dimensioni dell'immagine - tipicamente 8 x 8 px)

per quello queste operazioni si fanno nella fase di pre-processing e non on-the-fly: sono molto costose!

Storage size

- # of pixel: 628×1024
- # of bytes: $628 \times 1024 \times 3$



Maria Luisa Sapino (BDMM 2010)

immagine di $n \times m$: # di pixel nell'img = $n \times m$

tipicamente ciascun colore viene codificato in 3 bytes (formato RGB):
ciascun colore viene visto come composizione additiva di tre colori
RED GREEN BLUE che compogono il colore così come viene percepito
dall'occhio umano.

1 byte associato a RED (da 0 a 255), 1 a GREEN e 1 a BLUE => in totale sono necessari $n \times m \times 3$ bytes

servono tecniche per comprimere immagini:
COME RIDURRE LA QUANTITÀ DI MEMORIA?

1) cambiare sistema di codifica. ora è colore per ogni pixel, ma i dati presentati così possono contenere ridondanze e sprecare memoria: si può cercare di eliminare la ridondanza, non informazione che non è importante ma semplicemente il cui contenuto informativo può essere derivato da altri che ho già memorizzato -> compressione LOSSLESS: non fa perdere informazione

What to do???

- Change coding
 - Data have coding redundancies
 - Get rid of them
 - Lossless!
- Examples:
 - Static/dynamic huffman coding
 1. Count the character frequencies
 2. Find patterns that are repeated (less information!!)
 3. Replace them with shorter codewords
 - Arithmetic coding
 - LZW (Lempel-Ziv-Welch) coding

Maria Luisa Sapino (BDMM 2010)

18
8

es:

- * PER PATTERN RIPETUTI: posso comprimere informazione contigua uguale in coppia (#pixel contigui con stesso colore, RGB)
 - * CODIFICA DI HUFFMAN PER STRINGHE: associare a ciascun carattere codice numerico di lunghezza costante, oppure posso fare analisi statistica -> alcuni caratteri più frequenti di altri e quindi posso associare loro dei codici più corti, mentre associo i codici più lunghi ai caratteri meno frequenti.
 - * CODIFICA ARITMETICA: stringa -> numero
- * LZW: associano codice ai soli pattern effettivamente presenti nei dati che si devono memorizzare

es: vocabolario lingua italiana ha 21 lettere, quante stringhe di lunghezza 10 potrei generare? 10^{10} , ma in realtà le parole nella lingua italiana di lunghezza 10 sono molte meno: associo codice a ciascuna lettera dell'alfabeto e genero codice per ogni sottostringa che trovo nel testo, così genero codice solo per sequenze che esistono effettivamente nel linguaggio (es genero codice per "ACC" ma non per "ACZ")

What to do???

- Lossy compression
 - Image may contain details that human eye can not recognize
 - Use domain transformation
 - Convert images from spatial domain to frequency domain
 - DCT (discrete cosine transform)
 - DFT (discrete fourier transform)
 - DWT (discrete wavelet transform)
 - Get rid of the frequencies which do not contain information

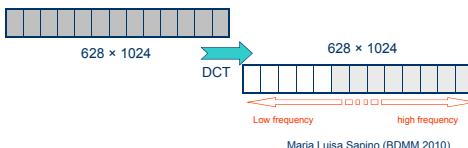
Maria Luisa Sapino (BDMM 2010)

18
9

a volte però non è sufficiente eliminare le ridondanze -> codifica LOSSY: fa perdere informazione "meno importante", cioè meno percepibile agli occhi dell'utente: trasferire il dato in un nuovo spazio in cui sia più facile isolare la parte di informazione che è più dannoso perdere. alcune conversioni (da dominio dell'ampiezza del segnale ad un dominio delle frequenze) fanno sì che nel nuovo dominio sia più facile isolare l'informazione utile; se la precisione è tale da non essere nemmeno percepibile dall'occhio umano, allora si può perdere dell'informazione senza degradare troppo la qualità del risultato

Domain transformations

- Given a signal of length "n" you get a sequence of "n" frequencies



Maria Luisa Sapino (BDMM 2010)

19
0

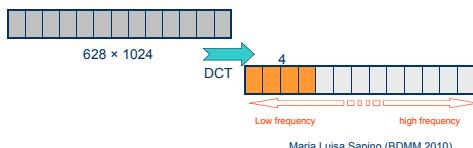
dato un segnale lungo n, attraverso la trasformazione si passa a rappresentazione dello stesso segnale e stessa lunghezza, però ora i valori in coda sono quelli "meno energetici", meno informativi: se li lascio cadere li perdo ma non faccio troppo danno dal punto di vista della qualità dell'immagine (è LOSSY).

...perché?

perché mentre nel primo dominio non c'è relazione tra i valori presenti nel dominio (non c'è semantica associata alla posizione), nel secondo invece ho stesso numero di coefficienti che sono associati a frequenze diverse, e le frequenze più basse sono all'inizio (sono in ordine crescente), frequenze più alte rappresentano il maggior grado di dettaglio e quindi sono quelle che posso lasciar cadere senza che siano percepite dall'occhio umano -> spostandomi e troncando via via verso sinistra faccio più danni, ovvero cancello parti importanti dell'immagine.

Domain transformations

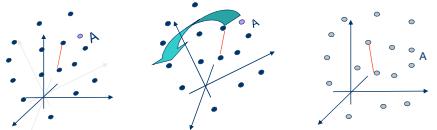
- Given a signal of length "n" you get a sequence of "n" frequencies



19
1

Domain transformations

- Given a signal of length "n" you get a sequence of "n" frequencies
 - preserves euclidean distance



Maria Luisa Sapino (BDMM 2010)

19
2

nei domini trasformati si preserva la distanza:
fondamentale nell'information retrieval perché
vogliamo che il retrieval sia equivalente nei due domini

Domain transformations

- Given a signal of length "n" you get a sequence of "n" frequencies
 - not sensitive to shifts!
 - take an image and shift it...both images will be mapped to the same point

Maria Luisa Sapino (BDMM 2010)

19
3

spostandoci a dominio di frequenza manteniamo proprietà importanti:

- * non è sensibile agli shift -> anche se shift l'immagine, la stessa immagine nel dominio trasformato viene ancora codificata nello stesso dominio di valori -> importante per interrogazioni in cui si cerca oggetto in un'img a prescindere da dove si trovi nell'img
(idem per rotazione)

Domain transformations

- Given a signal of length “n” you get a sequence of “n” frequencies
 - concentrates energy better

628 × 1024

DCT

These also have large values

Low frequency high frequency

Removing these will cause small errors!

Maria Luisa Sapino (BDMM 2010)

2D-DCT

Meet ‘Lena’: this picture is the most commonly used benchmark in image processing

Pixel Domain

Based on Ciriani et al. "Microdata Protection" 2007.

2D-DCT

Pixel Domain

Frequency Domain

Based on Ciriani et al. "Microdata Protection" 2007.

posso rappresentare immagine di Lena come sequenza di $n \times n$ coefficienti: f più bassa per $[0,0]$, spostandomi verso destra aumenta f in verticale (verso il basso aumenta f in orizzontale)

Lossy compression

Pixel Domain

70	70	100	70
85	100	96	79
100	85	116	79
136	69	87	200
161	70	87	200
161	123	147	133
146	147	175	100
156	146	189	70

Frequency Domain

-80	-40	89	-73
-135	-59	-26	6
47	-76	66	-3
-2	10	-18	0
-1	-9	-22	8
5	-20	28	-46
6	-20	37	-28
-5	-23	33	-30

Based on Ciriani et al., "Microdata Protection" 2007.

19
7

matrice descrive il quadratino dell'immagine

$F(u, v)$: il valore della matrice delle frequenze in base al valore della matrice di partenza (ampiezze)

la funzione con i COS calcola la conversione tra le due matrici.

matrice delle frequenze: immagine come somma di frequenze in cui quelle più alte sono quelle che portano più informazione

Lossy compression

Frequency Domain

-80	-40	89	-73
-135	-59	-26	6
47	-76	66	-3
-2	10	-18	0
-1	-9	-22	8
5	-20	28	-46
6	-20	37	-28
-5	-23	33	-30

Lossy Frequency Domain

-80	-44	90	-80
-132	-60	-28	0
42	-78	64	0
0	17	-22	0
0	0	0	0
0	0	0	0
0	0	0	0
0	0	0	0

Based on Ciriani et al., "Microdata Protection" 2007.

19
8

elimino dettagli nell'immagine

Lossy compression

Lossy Frequency Domain

-80	-44	90	-80
-132	-60	-28	0
42	-78	64	0
0	17	-22	0
0	0	0	0
0	0	0	0
0	0	0	0
0	0	0	0

Lossy Pixel Domain

70	60	106	94
85	101	85	75
98	99	92	102
132	53	111	180
173	57	114	207
164	123	131	135
141	159	169	73
150	141	195	79

Based on Ciriani et al., "Microdata Protection" 2007.

19
9

nella nuova immagine ho perso precisione

se a questo punto voglio tornare all'immagine iniziale, applicando DCT inversa otengo un'immagine diversa da quella di partenza, perché ho perso informazione

Lossy compression

matrice di errore
20 (differenza tra
0 matrici)

70 70 100 70 87 87 150 187	70 60 105 94 62 103 146 176
85 100 96 79 87 154 87 113	65 101 65 75 102 127 93 144
100 85 116 79 70 87 86 196	98 99 92 102 74 98 89 167
136 69 87 200 79 71 117 96	132 53 111 180 55 70 105 145
161 70 87 200 103 73 96 113	173 57 114 207 111 89 84 90
161 123 147 133 113 113 85 161	164 123 131 135 133 92 85 162
146 147 175 100 103 103 163 187	141 159 169 73 106 101 149 224
156 146 189 70 113 161 163 197	150 141 195 79 107 147 210 153

Based on Ciriani et al., "Microdata Protection" 2007.

PER COMPRIMERE si troncano i coefficienti con massima frequenza (cioè che portano più precisione): se faccio riconversione, ovvero parto da dominio delle frequenze e applico la funzione inversa, ottengo visualizzazione diversa da quella di partenza, cioè meno precisa -> posso calcolare la matrice differenza per quantificare l'errore

Meet "DCT bases": these are used in JPEG compression

2D-DCT

Pixel Domain Frequency Domain

20
1

Based on Ciriani et al., "Microdata Protection" 2007.

PROBLEMA: se eliminiamo frequenze più alte, questo ci porta a sottostimare le distanze => nel dominio compresso due immagini risultano più simili

questo è positivo o negativo?

* negativo perché gli oggetti sembrano più simili e quindi avrò più false-hits, ma anche positivo perché avrò meno misses se avessi allargato le distanze allora rischierai di perdere immagini simili (è sempre preferibile avere immagini simili in surplus e poi eliminarle con post-processing)

Under-estimation of distances

- If we remove some frequencies, this will cause underestimation of distances.
 - Why?
 - Why is this important?

628 × 1024 DCT 4

Low frequency high frequency

20
2

$\Delta > \Delta'$ Luisa Sapino (BDMM 2010)

...SO...

- In the design of a MIS, you want to minimize the number of image analysis operations to perform on the fly
 - pre-processing/ pre-analysis
 - indexing/clustering
 - semantic optimization
 - If
 - $(op1 \circ op2) = (op2 \circ op1)$ and
 - $\text{Cost}(op2 \circ op1) < \text{Cost}(op1 \circ op2)$

20
3

Maria Luisa Sapino (BDMM 2010)

Color

- Each pixel
 - 24 bits (3 bytes) of red, green, and blue
 - 2^{24} colors \approx 16 million

20
4

Maria Luisa Sapino (BDMM 2010)

What to do???

- Lossy compression
 - Image may contain details that human eye can not recognize
 - Color table:
 - reduce the number of colors to 256 (1 byte per pixel)
 - Cluster similar colors into a single bucket and assign a single color to the bucket
 - the set of buckets is called **color table**

20
5

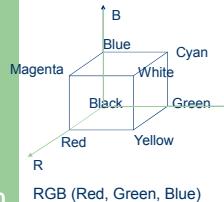
Maria Luisa Sapino (BDMM 2010)

altro modo di comprimere: utilizzare la **TAVOLOZZA DEI COLORI** sfruttando il fatto che l'occhio umano non è sensibile a tal punto da distinguere tutti i colori che possono essere rappresentati

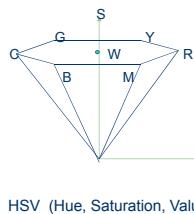
es: differenze su colori che differiscono per valore 1 su ciascuna componente R/G/B non sono percepite dall'occhio umano

quindi si possono raggruppare colori simili in un unico cluster, associarli ad un unico rappresentante, e ogni volta che nell'immagine di partenza compare un colore di quel gruppo, anziché rappresentare il colore preciso si usa il rappresentante del gruppo -> appiattisco la rappresentazione e perdo dettagli, ma riduco di molto lo spazio dei valori necessari per la rappresentazione

Alternative color spaces...



20
6
RGB (Red, Green, Blue)



HSV (Hue, Saturation, Value)
Maria Luisa Sapino (BDMM 2010)

esistono modelli alternativi, ma sono tutti equivalenti ovvero traducibili l'uno nell'altro facilmente:

* RGB: esistono 3 colori fondamentali (Red Green Blue) e gli altri sono definiti come combinazioni di questi tre colori

* HSV (Hue Saturation Value): rappresenta i colori come un cono in cui l'asse rappresenta la quantità di energia (da nero a bianco), angolo formato dal vettore che unisce l'asse alla superficie del cono identifica la tinta (hue), e la distanza dall'asse rappresenta la saturazione

sono due modelli equivalenti

Color models

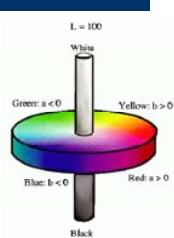
- RGB: describes colors in terms of the combinations of the intensities of Red, Green and Blue colors
- HSV
 - Hue: main color
 - Saturation: Amount of white
 - Value: Amount of energy

20
7
Maria Luisa Sapino (BDMM 2010)

in alcuni modelli la differenza nella codifica modella meglio la differenza nella percezione umana di colori diversi, es: CIELAB

Why different color models?

- Some color models reflect human perception better.
- Ex: CIELAB models the perceived differences in color and brightness



Maria Luisa Sapino (BDMM 2010)

YUV model

- YUV, a linear transformation from RGB
 - Y: luminance (amount of light) – grey scale
 - U: red - cyan
 - V: magenta-green

20
9

Maria Luisa Sapino (BDMM 2010)

trasformazione lineare di RGB dove si associano 3 bytes per ciascun colore, dove:

* Y è la luminanza -> quantità di luce (rappresentazione del colore in scala di grigio)

* U è la variazione di colore da rosso a ciano

* V è la variazione di colore da magenta a verde

YUV (ex. PAL television system)

21
0

Y=0.299R + 0.587 G + 0.114 B
U= 0.492 (B-Y)
V= 0.877 (R-Y)

Maria Luisa Sapino (BDMM 2010)

quale vantaggio di YUV rispetto a RGB?
ci permette di avere immediatamente

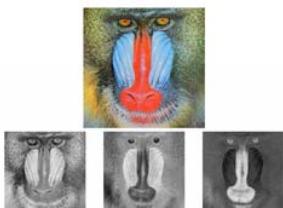
rappresentazione in bianco e nero, e poi si può passare alla versione a colori aggiungendo i due canali U e V

YUV transforms the space

21

Maria Luisa Sapino (BDMM 2010)

Y is more important than U and V



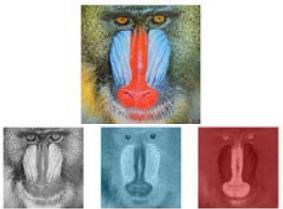
21
2

Li & Drew © Prentice Hall 2003

Maria Luisa Sapino (BDMM 2010)

Y è più importante rispetto a U e V: è la componente che porta più dettaglio e quindi non va "appiattita" (compressa)

Y is more important than U and V



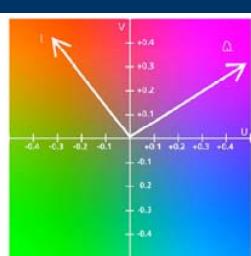
21
3

Li & Drew © Prentice Hall 2003

Maria Luisa Sapino (BDMM 2010)

in realtà in questo spazio di colori U e V sono equivalenti dal punto di vista della percezione, ma la stessa cosa non vale per le diagonali: se passo a questo nuovo spazio con le due diagonali come assi ottengo sistema YIQ -> l'occhio umano è più sensibile ai colori nella direzione dal rosso al blu, quindi il candidato ad essere compresso è Q

...not enough (YIQ)...



21
4

Maria Luisa Sapino (BDMM 2010)

..not enough (YIQ)..

Li & Drew ©Prentice Hall 2003

Human eye is more sensitive to the Orange-Blue range than the Purple-Green range.

Maria Luisa Sapino (BDMM 2010)

21
5

se definisco come assi principali I e Q, nel nuovo sistema osservo che l'occhio umano è più sensibile alle variazioni che vanno dall'arancione al blu

la componente candidata ad essere soppressa è quindi Q

TAVOLOZZA COLORI: rappresentare colori tramite cluster -> es: 2 cluster con cui rappresento colori in cui il verde è <= della media e verde è > della media

What to do???

- Lossy compression
 - Image may contain details that human eye can not recognize
 - Change the color representation (don't use RGB!)
 - human eye is not sensitive to color anyway
 - Human eye is more sensitive to contrast!
 - Use a representation which increases the contrast and compress the color element
 - Use YRB (Luminance, Red, Blue)
 - Quantization:
 - 1 byte for Luminance
 - ½ byte for Red
 - ½ byte for Blue

Maria Luisa Sapino (BDMM 2010)

21
6

Vogliamo estrarre il surrogato di un'immagine:
primo approccio è realizzare istogramma dei colori -> funzione che ad ogni colore presente in un'immagine associa informazione sulla frequenza del colore nell'immagine.
--> istogramma è SURROGATO dell'immagine

Color histograms

Color Category	Frequency
Dark Gray	3.4
Light Gray	18.7
Black	1.6
Blue	21.3
Green	9.3
Yellow	11.1
Red	11.1
Orange	9

Maria Luisa Sapino (BDMM 2010)
Courtesy of Misha Pavel, OGI

21
7

Problems with histograms



Histogram: {green:4, purple:2, red:3}

- Are these similar???

21
8

Maria Luisa Sapino (BDMM 2010)

istogramma è insensibile alla località dei colori, ovvero
"dove si trova il colore?"

quindi nelle query del tipo "immagini simili" due immagini
risultano simili se gli istogrammi sono simili,
indipendentemente dalla posizione dei colori

Problems with histograms



Histogram: {green:4, purple:2, red:3}

- Are these similar???

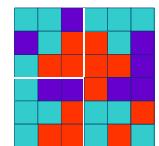
- Color associations????:
 - blue is similar to purple
 - yellow is similar to orange

21
9

Maria Luisa Sapino (BDMM 2010)

associazione dei colori non viene più mantenuta

Color locality??



22
0

Maria Luisa Sapino (BDMM 2010)

posso suddividere l'immagine in varie porzioni e calcolare
l'istogramma dei colori per ogni porzione, così mantengo
un po' di informazione sulla localizzazione dei colori:
più porzioni faccio e meglio gestisco la località dei colori!

Comparison of color histograms

- Euclidean distance

$$\sqrt{(b_1-b_2)^2 + (g_1-g_2)^2 + (p_1-p_2)^2 + (r_1-r_2)^2 + \dots}$$

- Intersection similarity

$$\min(b_1, b_2) + \min(g_1, g_2) + \min(p_1, p_2) + \min(r_1, r_2) + \dots$$

$$b_2 + g_2 + p_2 + r_2 + \dots$$

Maria Luisa Sapino (BDMM 2010)

22
1

come calcolo la somiglianza tra istogrammi?

- distanza euclidea: calcola radice quadrata della somma delle differenze dei valori corrispondenti
- somiglianza rispetto all'intersezione: quale percentuale di istogramma è in comune tra i due istogrammi? = in quale percentuale i due istogrammi si intersecano?

Complete Euclidean Distance

- Let x and y be two histogram vectors, each of length n

$$d^2 = \sum_{i=1..n} \sum_{j=1..n} a_{ij} (x_i - y_j)^2$$

- a_{ij} = cross talk factor between i -th and j -th color

$$\text{No cross-talk} \rightarrow \begin{cases} a_{ij} = 1 & \text{if } i = j \\ a_{ij} = 0 & \text{otherwise} \end{cases}$$

22
2

Maria Luisa Sapino (BDMM 2010)

le differenze non pesano tutte allo stesso modo:

es: differenza tra viola e porpora è minore di differenza tra blu e arancione per tenere conto di questo si usa un correttivo alla distanza euclidea: anziché limitarsi a fare sommatoria della differenza dei quadrati degli stessi colori in due diversi istogrammi, qui si combina ciascun colore presente in un istogramma con ciascun colore presente nell'altro istogramma, si calcola la distanza tra i singoli colori combinati con un fattore che indica quando sono simili i due valori (es: x è istogramma1, y è istogramma2, i è colore rosso, j è colore verde, aij fattore di cross-talk che indica somiglianza tra rosso e verde)

|

|

V

è strategia molto costosa!

Quadratic distance bounding

- Use average color of an image

$$R_{avg} = (1/N) \sum_{i=1..N} R(p_i) \quad G_{avg} = (1/N) \sum_{i=1..N} G(p_i) \\ B_{avg} = (1/N) \sum_{i=1..N} B(p_i)$$

$$\underline{x} = (R_{avg}, G_{avg}, B_{avg})^T$$

$$d_{avg}^2 (x, y) = (x - y)^T (x - y)$$

$$= (R_{avgx} - R_{avgy})^2 + (G_{avgx} - G_{avgy})^2 + (B_{avgx} - B_{avgy})^2$$

22
3

Maria Luisa Sapino (BDMM 2010)

quindi si usa un bounding quadratico per la distanza:
per confrontare due immagini calcolo distanza euclidea tra le loro rappresentazioni di colore MEDIO.

(D_{avg}^2 = distanza quadratica media)
distanza media è più economica

Quadratic distance bounding

$$d_{avg}^2(x,y) \leq c \cdot d_{hist}^2(x,y)$$

22
4

Maria Luisa Sapino (BDMM 2010)

la distanza quadratica media è sempre \leq di una certa costante C (che può essere calcolata) per la distanza esatta -> mette in relazione la stima con la distanza esatta.

le immagini vengono avvicinate!

serve nelle queries sulla distanza tra immagini in cui si specifica la distanza massima che deve esserci tra le due immagini:

es: immagini che distano al più R/C -> vengono restituiti possibili false hits ma nessun miss, perché le immagini risultano più vicine

Quadratic distance bounding

$$d_{avg}^2(x,y) \leq c \cdot d_{hist}^2(x,y)$$

Why is this good??

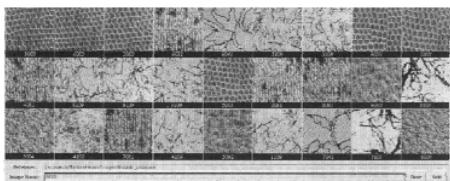
economico
e nessun miss

22
5

Maria Luisa Sapino (BDMM 2010)

vedi 05.txt per dettagli su texture e forme

Texture



22
6

Maria Luisa Sapino (BDMM 2010)
Courtesy of Misha Pavel, OGI

se considero matrice di texture

$$\begin{bmatrix} \square & \times & \times \end{bmatrix}$$

$$\begin{bmatrix} \times & \square & \times \end{bmatrix}$$

$$\begin{bmatrix} \times & \times & \square \end{bmatrix}$$

prima calcolo $dx = 100 \times (-1) + 0 \times 0 + 1 \times 0$ ecc.ecc.
e stessa cosa per dy ,

dopodiché calcolo $\sqrt{dx^2 + dy^2}$ e ottengo la variazione di intensità, $\arctan(dy / dx)$ e ottengo la direzione

non è sufficiente considerare i singoli pixel per valutare la texture, in quanto non è una proprietà locale ai singoli pixel ma è una proprietà globale di gruppi di pixel:
variazione di intensità e direzionalità del segnale luminoso nell'immagine

(N.B. poiché è "variazione" è necessario confronto tra almeno 2 pixel!)

ad ogni pixel si associa coppia [intensità, direzione] (angolo)

si calcola la variazione dell'intensità di luminosità, poi si filtra togliendo i pixel di "disturbo" che sono sotto una certa soglia di luminosità (percentuale o threshold)

es: per matrice 3 x 3 si ha

$$dx$$

$$\begin{bmatrix} -1 & 0 & 1 \end{bmatrix}$$

$$\begin{bmatrix} -2 & 0 & 2 \end{bmatrix}$$

$$\begin{bmatrix} -1 & 0 & 1 \end{bmatrix}$$

$$dy$$

$$\begin{bmatrix} 1 & 2 & 1 \end{bmatrix}$$

$$\begin{bmatrix} 0 & 0 & 0 \end{bmatrix}$$

$$\begin{bmatrix} -1 & -2 & -1 \end{bmatrix}$$

fare CLUSTERING: raggruppare in cluster (= gruppi omogenei) oggetti simili (N.B. bisogna prima definire la somiglianza tra oggetti)

ci sono più modi per clusterizzare:

es: 3) a partire da un pixel si fa crescere la regione attorno fino a che non si può più allargare

Texture

- Primitives:
 - Grey level
 - Shape
 - Homogeneity
 - Smoothness
 - Finess
 - Coarseness
 - Granularity
 - Regularity
 - Linearity
 - Directionality

22
7

Maria Luisa Sapino (BDMM 2010)

CHAINING:

+ è facile

+ è intuitivo

+ non è sensibile a traslazione (= spostamento della forma lungo uno degli assi)

ma

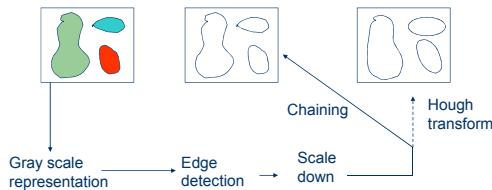
- è sensibile al punto in cui si decide di iniziare, ma si può scegliere un punto di default (es: punto più in basso a sx)

- più grave -> è sensibile alla rotazione perché cambiano le direzioni delle variazioni, quindi non permette di riconoscere due immagini uguali ma ruotate in modo diverso

- è sensibile alla scalatura dell'immagine: per stessa immagine su scala diversa ho stringhe diverse

la codifica è precisa a livello di pixel, quindi immagini abbastanza simili possono essere codificate da stringhe di lunghezza diversa -> vogliamo stringhe di lunghezza fissa -> ricorriamo a tecniche di interpolazione rappresentando forme chiuse con un certo numero di segmenti

Edges



22
8

Maria Luisa Sapino (BDMM 2010)

per ora non ho dato nessuna descrizione semantica (es: cerchio, ellisse), ho solo parlato dei pixel -> ora vogliamo dare una rappresentazione intensionale -> un modo è usare la TRASFORMATA DI HOUGH

Hough Transform

- a technique for isolating features of a particular shape within an image.
- computes a global description of a feature given (possibly noisy) local measurements.
- Most commonly used for the detection of regular curves such as lines, circles, ellipses
- Generalized Hough transform in applications where simple analytic description of features is not possible

22
9

Maria Luisa Sapino (BDMM 2010)

tecnica che permette di associare a forma nell'immagine una sua descrizione analitica (es: interpolazione di punti -> retta)

TRASFORMATA: descrizione analitica della curva viene data in un nuovo spazio trasformato che non è lo stesso di partenza

es: interpolazione punti su retta $y = ax + m$: ho i punti da interpolare e voglio trovare le incognite a e m -> mi basterebbero due equazioni, ma in questo caso c'è rumore e quindi non si può!

un modo per farlo è usare un meccanismo di VOTING: si cerca di indovinare i valori di a e m più plausibili, ovvero supportati dal maggior numero di punti -> costruisco una matrice in cui ho su una dimensione i possibili valori di a e sull'altra i possibili valori di m

* tutte le coppie (a, m) supportate dalla prima equazione $y_1 = a x_1 + m_1$ vengono poste a 1, e così viene fatto per tutti i punti

* alla fine vince la coppia più votata

se si identificano più coppie con valori alti, allora la retta non era una sola e quindi le equazioni trovate rappresentano diverse rette da considerare

si può anche fare sulle altre forme usando l'equazione relativa e costruendo altra matrice con i parametri necessari

Shape

- Segmentation/tiling (based on color or illumination)

- A segmentation of image I is a set $\{R_1, \dots, R_n\}$ of regions such that:
 - $R_1 \cup \dots \cup R_n = I$
 - $R_i \cap R_j = \emptyset$ for all $1 \leq i, j \leq n$
 - $H(R_i) = \text{true}$ for all $1 \leq i \leq n$
 - If R_i, R_j are adjacent, then $H(R_i \cup R_j) = \text{false}$.

23
0

Maria Luisa Sapino (BDMM 2010)

partizionare l'immagine in regioni:

-unione delle regioni dà l'immagine di partenza

-le regioni sono separate

-ciascuna regione soddisfa un predicato di omogeneità

(es: stesso colore, stessa tessitura, ecc.)

-regioni adiacenti non devono essere omogenee

dobbiamo riconoscere che i segmenti sono disposti in un certo modo per riconoscere una certa FORMA

Shape

- Segmentation/tiling (based on color or illumination)

- Hough transform of edges
 - Global features
 - Roundness
 - Aspect ratio (ratio between the width of the image and the height of the image)
 - Major axis orientation
 - Local features
 - Size and orientation of consecutive boundary segments

23
1

Maria Luisa Sapino (BDMM 2010)

segmentazione è basata su colore o illuminazione e permette di estrarre due tipi di proprietà:

* proprietà GLOBALE -> caratterizza nella sua globalità il segmento che descrive (es: per segmento chiuso il grado di rotondità, se è = 1 è un cerchio, oppure rapporto tra allungamento in verticale/orizzontale, oppure ancora orientamento dell'asse principale)

* proprietà LOCALE: relativa a frammento del segmento (es: spezzata che identifica regione chiusa, ciascun segmento con propria lunghezza e inclinazione rappresenta una feature locale)

Spatial queries

NO

23
2

(BDMM 2010)

Spatial queries

-fingerprint detection
 - spatial distribution
 - positions of features relative to each other
 - ..their angles
 - ..their distances from each other

http://www.sphnxtch.com.sg/fingerfeatures.jpg
http://www.informatik.htw-dresden.de/~ive/Belege/Poetzsch/
Maria Luisa Sapino (BDMM 2010)

Spatial relationships

- How do we represent the spatial relationships among these objects?
 - to the left, to the right
A is to the left of B
 - above of, below of
A is above/below B?

Minimum Bounding Rectangles (MBR)
Maria Luisa Sapino (BDMM 2010)

Comparing the spatial relationships

NO

- One solution is to assume that they are equal (however, this solution is not transitive)
- Second solution is to divide A and B into smaller pieces, and describe the relationships of the pieces
- Third solution is to use INTERVALS

Maria Luisa Sapino (BDMM 2010)

informazioni di carattere spaziale, ovvero informazioni su posizione reciproca di oggetti nell'immagine: come misurare e quantificare posizione reciproca?

* in alcune applicazioni è importante la posizione reciproca tra elementi dell'oggetto per i confronti (es: impronte digitali)

come rappresentiamo le relazioni spaziali tra gli oggetti?
c'è ambiguità:

* sicuramente A è a sinistra di B, perché TUTTI i punti di A sono a sinistra di TUTTI i punti di B.

* A è sopra/sotto B?

dipende da quale semantica associamo alla relazione "stare sopra" o "stare sotto" (in assoluto A non sta né sopra né sotto B)

il modo più semplice per confrontare è regolarizzare gli oggetti: racchiudiamo ciascuna forma in un rettangolo (MBR = Minimum Bounding Rectangle) i cui lati sono tangenti al perimetro della figura, e poi confrontiamo i rettangoli (si può fare perché la superficie della forma occupa in percentuale la maggior parte del rettangolo)

Minimum bounded rectangle

23
6

- Too simple..

Maria Luisa Sapino (BDMM 2010)

-) può portare a risultati scorretti, in caso di query sul grado di sovrapposizione degli oggetti, in quanto i rettangoli possono sovrapporsi anche se le figure all'interno non si sovrappongono!

Plane sweep

23
7

- Each vertex is an event point
 - Store the status of sweeplines for each event point
- Good for containment and overlap queries

Maria Luisa Sapino (BDMM 2010)

una soluzione è il PLANE SWEEP: non mi limito a circoscrivere un rettangolo, ma cerco punti di intersezione tra piano verticale che faccio scorrere lungo l'asse delle X e catturo punti di tangenza a intervalli regolari tra il piano e la figura: così ottengo un poligono la cui forma è più fedele all'oggetto rappresentato

questa strategia è più adatta per le query di sovrapposizione!

Spatial orientation graph

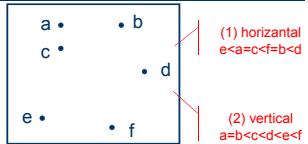
NO

23
8

- Weights of the edges are the slope of the corresponding line

Maria Luisa Sapino (BDMM 2010)

2D string



= ← at the same location
< ← to the left (below)
: ← in the same set as

23
9

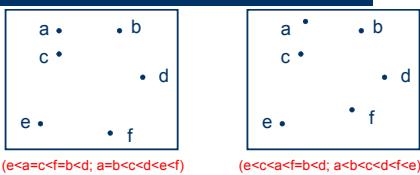
Maria Luisa Sapino (BDMM 2010)

come rappresentare le relazioni spaziali tra i punti?

- posso fare collezione di coppie (a, b) con vari significati (es: a è a sinistra di b, a è alla stessa altezza di b, ecc...)
- altro modo (equivalente a planesweep): lettura dei punti da sx verso dx e poi dall'alto verso il basso e si rappresentano le stringhe ottenute dalla lettura (si deve però stabilire una relazione d'ordine tra punti equivalenti, es: a c):
le due stringhe rappresentano un surrogato della rappresentazione bidimensionale dell'immagine => ora ho rappresentazione unidimensionale come stringhe

N.B. operatore : serve per raggruppare punti, utile per le query di posizione reciproca tra oggetti

Compare space = compare string



24
0

Maria Luisa Sapino (BDMM 2010)

si confrontano gli oggetti tramite il confronto tra le stringhe nello spazio bidimensionale

PROBLEMA: se ruoto l'immagine cambia la stringa che la rappresenta perchè cambia l'ordine con cui il piano incontra i punti della figura

How do we compare strings??

(e<a=c<f=b<d; a=b<c<d<e=f) vs. (e<c<a<f=b<d; a<b<c<d<e)

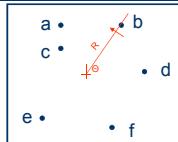
- Edit distance:
 - "table" vs. "cable": 1 (replace "t" with "c")
 - "table" vs. "bale": 2 (delete "l"; swap "a" and "b")

24
1

Maria Luisa Sapino (BDMM 2010)

distanza tra stringhe si misura con l'edit distance: misura il costo minimo della trasformazione di una stringa in un'altra tramite operazioni elementari (inserimento, cancellazione, sostituzione): applichiamo edit distance tra le due stringhe = due oggetti distano quanto il costo di trasformazione tra le rispettive stringhe

OR string



24
2

Maria Luisa Sapino (BDMM 2010)

per risolvere problema delle stringhe sulla rotazione (vedi sopra) si introducono delle varianti: si parte dal centro dell'immagine e si fa "spazzare" l'immagine dal centro come le lancette di un orologio (si calcola anche la lunghezza del raggio dal punto al centro per gestire casi di uguaglianza)

è dipendente da quale "ora" inizio, ma si può risolvere allineando i due punti di partenza sulle due immagini: scelgo un punto che sia presente in entrambe le immagini che devo confrontare (non è necessario che siano allineati) e faccio partire le lancette da lì

N.B. è ancora sensibile alla scalatura

come scegliamo i punti da associare all'immagine?

1) vertici della bounding box

2) centroide dell'immagine -> più facile da utilizzare, utile per query del tipo "o1 è dominantemente a dx di o2?", non risulta più grande di quanto sia l'oggetto (a differenza della bounding box)

String representations of space

- 2D string
 - Use centroids (points)
 - 3 operators
- 2D Estring
 - Use intervals
 - Compare intervals
 - More operators
 - Contain, meet, begin, end, overlap, equal, less than, overlap, inverse

24
3

Maria Luisa Sapino (BDMM 2010)

METODI ALTERNATIVI:

- non si associa direttamente un punto all'immagine ma si associano degli intervalli, però richiede più operatori

due approcci:

1) associare a ciascuna immagine il suo punto più a sx e il suo punto più a dx (idem per sopra e sotto): richiede presenza di operatore OVERLAP perché intervalli possono intersecarsi

String representations of space

- 2D Gstring
 - Divide objects into smaller objects
 - Then use 2D strings
- 2D Cstring
 - Less partitions, more operators (intervals)
- 2D Bstrings
 - Use end points, less operators

24
4

Maria Luisa Sapino (BDMM 2010)

2) dividere oggetti in oggetti più piccoli in modo da non avere contenimento di intervalli di immagini diverse l'uno nell'altro, ma solo relazioni di posizionamento reciproco (non serve OVERLAP)



NO

Audio

- SPEECH
 - Speech coding
 - Speech recognition
 - Voice recognition
 - Word recognition
 - Speech to text conversion
 - Speech synthesis
 - Access for blind individuals
 - Language to language conversion
- SOUND
 - Sound analysis
 - Extraction of the characteristics of a particular sound
 - Sound synthesis
 - Creation of a sound with particular characteristics

24

5

Audio vs. Image → temporal information

Maria Luisa Sapino (BDWIM 2010)

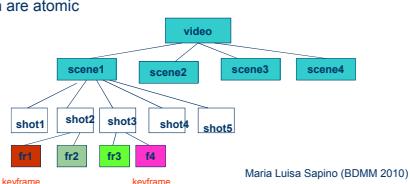
NO

Audio

- SPEECH
 - Speech coding
 - Speech recognition
 - Voice recognition
 - Word recognition
 - Speech to text conversion
 - Speech synthesis
 - Access for blind individuals
 - Language to language conversion
 - SOUND
 - Sound analysis
 - Extraction of the characteristics of a particular sound
 - Sound synthesis
 - Creation of a sound with particular characteristics

Audio vs. Image → temporal information

Maria Luisa Sapino (BDMM 2010)



24
6

Video (temporal structure!!!!)

- Time (as in audio) + objects and motion
 - Frame by frame representation is too costly (30f per second)
 - Shot detection (video segmentation) gives a set of frames which are atomic

VIDEO: sequenza di scene

SCENA: sequenza di shots

SHOTS: sequenze di fotogrammi (frame) che corrispondono allo stesso punto di vista della telecamera (quando cambia il punto di vista cambia lo shot)

scena: ha significato semantico -> sequenza di shot omogenei

di solito per le query si usano frames rappresentanti di ciascuno shot, non tutti i frames del video

per analizzare il video occorre

- Camera motion
 - Zooming (varying the focus distance)
 - Tilting (down/up - camera vertical rotation)
 - Panning (right/left - camera horizontal rotation)
 - Tracking (horizontal transverse movement)
 - Booming (vertical transverse movement)
 - Dollying (toward/away – horizontal lateral movement)
 - Object motion
 - Scene change
 - Problem: if you don't know about camera motion ahead in time, then it is harder to distinguish between object motion and camera motion

Maria Luisa Sapino (BDMM 2010)

- riconoscere opportunamente il movimento della telecamera e gli spostamenti reciproci degli oggetti
 - + ZOOMING: mettere a fuoco
 - + TILTING: rotazione lungo l'asse verticale
 - + PANNING: rotazione lungo l'asse orizzontale
 - + TRACKING: movimento trasverso in orizzontale
 - + BOOMING: movimento trasverso in verticale
 - + DOLLYING: movimento orizzontale laterale

PROBLEMA:
capire se è la telecamera o l'oggetto a muoversi!

NO

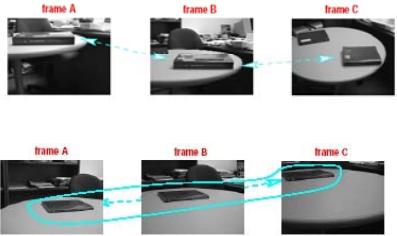
Motion analysis

- Approach
 - Look at the global motion to determine the camera motion (estimate)
 - Subtract the estimated camera motion
 - Result will be the object motion
- Object trajectory
 - B-spline
 - Chain code
 - Differential chain code

24
8

Maria Luisa Sapino (BDMM 2010)

Object tracking



frame A frame B frame C

frame A frame B frame C

24
9

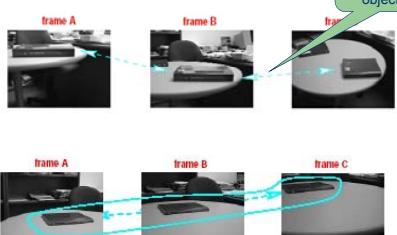
Maria Luisa Sapino (BDMM 2010)

PROBLEMA del RICONOSCIMENTO degli OGGETTI:
se si sposta la telecamera cambia anche la forma dell'oggetto
 (perché cambia la prospettiva!)

per riconoscere lo spostamento di un oggetto bisogna prima capire
 che due forme in diversi frame vicini sono lo stesso oggetto
 è un problema molto difficile (non abbiamo visto nel dettaglio)

NO

Object identity



frame A frame B frame C

frame A frame B frame C

25
0

Maria Luisa Sapino (BDMM 2010)

Time

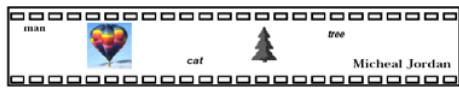
- What is time?
- How do we represent time?
- How do we represent actions and events?
- How do we ask queries about time?

25
1

Maria Luisa Sapino (BDMM 2010)

Maria Luisa Sapino (BDMM 2010)

Video queries



Object queries

Maria Luisa Sapino (BDMM 2010)

Video queries

Maria Luisa Sapino (BDMM 2010)

distanza temporale che intercorre tra due eventi:
voglio poter risolvere query del tipo "trovare partite di calcio in cui
due goal sono stati fatti a meno di 2 minuti di distanza" l'uno
dall'altro

- tempo è associato direttamente al numero di frame (tanto se sappiamo che proiettiamo 24 frames/se possiamo facilmente risalire al tempo)

- come rappresentiamo il tempo
 - come rappresentiamo azioni/eventi?
 - come facciamo query sul tempo?

diversi tipi di query temporali con complessità crescente:

si richiede che un certo numero di oggetti siano presenti nel video in una certa sequenza temporale:

tipo di query temporale più semplice, è sufficiente identificare gli oggetti nelle immagini relative ai frame

query più complicata: entro nel merito della struttura del singolo frame = ogni frame deve contenere particolari oggetti con particolari relazioni

26

Video queries

Simple action queries

25
4

Maria Luisa Sapino (BDMM 2010)

AZIONI SEMPLICI: riconoscere eventi in base a sequenza di frame in un certo ordine (es: goal in una partita richiede palla in porta, poi gente che esulta, ecc...): posso cercare video in cui sono presenti certe scene (= sequenze di fotogrammi)

posso mettere anche vincoli su distanza massima di frame successivi

Video queries

Composite action queries

25
5

Maria Luisa Sapino (BDMM 2010)

AZIONI COMPLESSE: nell'es. mongolfiera passa oltre alla nuvola a distanza massima di 8 frames

abbiamo espresso visivamente come un utente può fare una query

Situation calculus

NO

- World is represented as a set of situations, each describing the world at a single time instant
 - state-space model
- Action: a function from one situation to another
 - prerequisites
 - consequences
- Challenges:
 - simultaneous actions
 - what if there is an action but no change in state??

25
6

Maria Luisa Sapino (BDMM 2010)

NO

NO

What do we model?

- Past, present, and future
 - Static aspects:
 - Properties: A owns a car
 - Dynamic aspects:
 - Occurrences: A uses a car
 - Temporal logic
 - instants?
 - intervals?

25
7

Maria Luisa Sapino (BDMM 2010)

Time Models

25
8

Multilevel Scales (PPMM 0010)

come rappresentiamo il tempo?

- 1) per gli istanti esistono solo 3 relazioni, ovvero
[=], [<] e [>]

Time Models

- **Instants**
 - 3 instant based relations: $<$, $=$, $>$
- **Intervals**
 - 13 interval based relations

25
9

Maria Luisa Soares (PBMMM 2010)

2) per gli intervalli temporali (rappresentano la durata)
si hanno 13 relazioni, di cui 6 hanno ciascuno un duale
e l'EQUAL è simmetrico (non ha duale)

Intervals...

A before B		A overlaps B	
A after B		A overlapped by B	
A starts B		A meets B	
A started by B		A met by B	
A ends B		A equal B	
26 0		A during B	
		A containing B	

Maria Luisa Sapino (BDMM 2010)

13 relazioni per intervalli temporali

What else?

- occur (event, Interval)
- occurring (process, Interval)
- causation (one event causing another event)
- actions (agents/objects cause occurrences)
 - intentional actions
- beliefs
- plans etc.

26
1

Maria Luisa Sapino (BDMM 2010)

posso combinare asserzioni sulle azioni per ottenere altre asserzioni

Temporal Logic; Axioms

- $\text{before}(I_1, I_2) \text{ and } \text{before}(I_2, I_3) \rightarrow \text{before}(I_1, I_3)$
- $\text{meets}(I_1, I_2) \text{ and } \text{during}(I_2, I_3) \rightarrow \text{overlaps}(I_1, I_3) \text{ or } \text{during}(I_1, I_3) \text{ or } \text{meets}(I_1, I_3)$

26
2

Maria Luisa Sapino (BDMM 2010)

Definitions

- before(I1,I2) and before(I2,I3) -> before(I1,I3)
- meets(I1,I2) and during(I2,I3) ->
overlaps(I1,I3) or during(I1,I3) or meets(I1,I3)
- in(I1,I2) <-> during(I1,I2) or starts(I1,I2) or finishes(I1,I2)

26
3

Maria Luisa Sapino (BDMM 2010)

posso derivare relazioni che non erano presenti nell'algebra, dare un nome a certe relazioni derivate (es: IN è valido se è valida una delle tre asserzioni che la definiscono: DURING, START o FINISHES)

Rules

- before(I1,I2) and before(I2,I3) -> before(I1,I3)
- meets(I1,I2) and during(I2,I3) ->
overlaps(I1,I3) or during(I1,I3) or meets(I1,I3)
- in(I1,I2) <-> during(I1,I2) or starts(I1,I2) or finishes(I1,I2)
- holds(p,I) <-> forall i (in(i,I) -> holds(p,i))
- holds(and(p,q),I) <-> holds(p,I) and holds(q,I)
- holds(not(p),I) <-> forall i (in(i,I) -> not holds(p,i))

26
4

Maria Luisa Sapino (BDMM 2010)

Properties

NO

- The set of intervals over which a property holds is closed under the "in" relation
- Example: "I have a car"

26
5

Maria Luisa Sapino (BDMM 2010)

NO

Processes

- Example: "I am walking"
- May occur in a subinterval
- ..but, does not need to occur in all subintervals!

26
6

Maria Luisa Sapino (BDMM 2010)

NO

Events

- Event A occurs over interval I
 - No subinterval of I has event A occurring!!!!
- Example: "I walked to the store"

26
7

Maria Luisa Sapino (BDMM 2010)

Interval Algebra

- Predicates: 13 relationships between intervals
- Operators: and, or
- Result: all relationships that hold between every pair of intervals

26
8

Maria Luisa Sapino (BDMM 2010)

Interval Algebra

- Predicates: 13 relationships between intervals
- Operators: and, or
- Result: all relationships that hold between every pair of intervals
- overlaps(a,b) and starts(a,b) and meets(b,c)
 - what else can you deduce from these facts?????

26
9

a before c, b containing A

Maria Luisa Sapino (BDMM 2010)

insieme di relazioni che intercorrono tra coppie di eventi di cui conosciamo qualcosa sulla reciproca relazione temporale: per capire quali sono i possibili casi che possono soddisfare questa specifica, si usa un grafo (vedi sotto)

Minimal Labeling Problem

- NP-hard
- Special cases: if there are no "or"s, this can be solved in $O(n^3)$ time
 - constraint propagation algorithm

27
0

Maria Luisa Sapino (BDMM 2010)

si costruisce un grafo in cui

- * vertice = evento

* arco = relazione tra eventi

archi inizialmente sono etichettati con tutte le possibili relazioni tra i due vertici:
risolvere il problema equivale ad assegnare ad ogni arco il numero minimo di etichette tale che siano soddisfatte le specifiche

è problema NP-hard!

si semplifica se la formula che descrive lo scenario non contiene disgiunzioni

Point algebra

- Use "instants" instead of "intervals"
- Use "before", "equal", and "after" instead of the 13 interval relationships
- (similar to situation calculus)

27
1

Maria Luisa Sapino (BDMM 2010)

* interval algebra (visto ora)
* point algebra

POINT ALGEBRA usa istanti, quindi ha solo 3 operatori:
[=], [<] e [>]

Instants vs intervals

- Each interval can be described by two instants **st** and **et**

27
2

Maria Luisa Sapino (BDMM 2010)

i due approcci non sono equivalenti:
ci sono alcuni casi che non si possono catturare con i
punti, ma solo con gli intervalli

Instants vs intervals

- Each interval can be described by two instants **st** and **et**
- If there are no disjunctions, interval algebra = instance algebra
- $\text{during}(A,B) =$
 $\text{st}(A) > \text{st}(B) \text{ and } \text{et}(A) < \text{et}(B) \text{ and }$
 $\text{st}(A) < \text{et}(A) \text{ and } \text{st}(B) < \text{st}(B)$

a inizia dopo che B è iniziato e finisce
prima che B sia finito

27
3

Maria Luisa Sapino (BDMM 2010)

Instants vs intervals

- Each interval can be described by two instants **st** and **et**
- Are **instant based model** and **interval based model** easily interchangeable?
 - NO!
- Example:
 $(a \text{ before } b) \text{ or } (a \text{ after } b)$
 \neq
 $(\text{et}_a < \text{st}_b) \text{ or } (\text{et}_b < \text{st}_a)$

27
4

Maria Luisa Sapino (BDMM 2010)

Instants vs intervals

- Each interval can be described by two instants **st** and **et**
- Are instant based model and interval based model easily interchangeable?
– NO!
- Example: $(a \text{ before } b) \text{ or } (a \text{ after } b)$
 \neq
 $(et_a < st_b) \text{ or } (et_b < st_a)$

27
5

Maria Luisa Sapino (BDMM 2010)

altro approccio utile per punti di sincronizzazione tra oggetti: RETI DI PETRI -> modello per rappresentazione della concorrenza -> sistema viene modellato con un grafo bipartito, ossia in cui i vertici sono di due tipi: posti [O] e transizioni [|] -> nessun arco collega due vertici dello stesso tipo -> rappresenta sistemi dinamici in evoluzione

vertice: POSTO

un posto sta sempre tra due transizioni e una transizione sta tra due posti

POSTO -> stato del sistema

TRANSIZIONE -> condizione che si può verificare e far passare da uno stato all'altro

ogni transizione può essere attivata quando in tutti i posti in input è presente il numero di TOKEN necessario

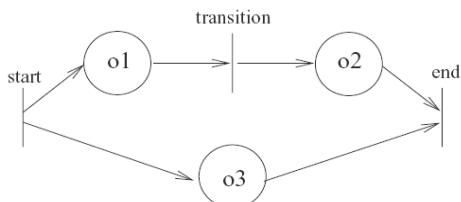
===== perché ci interessa?? =====

* ciascuno stato rappresenta un OGGETTO multimediale presente nel sistema (es: immagine, scene di un video, ecc...)

* nelle OCPN ciascuna transizione ha un unico oggetto in input e uno in output

* lo scatto della transizione non avviene indeterministicamente, ma la permanenza di un token in un oggetto è associata ad una durata -> transizione = momento di sincronizzazione: ad es. posso rappresentare un filmato o1 seguito da altro filmato o2 entrambi con colonna sonora o3 (transizioni sono momento di sincronizzazione)

Object Composition Petri Nets



Maria Luisa Sapino (BDMM 2010)

27
6

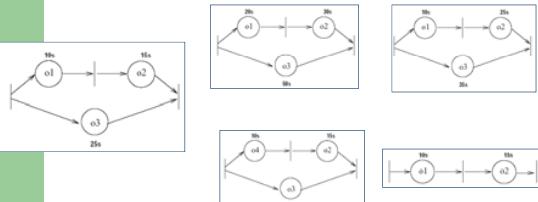
- posso studiare raggiungibilità di certi stati

per calcolare la somiglianza devo fare un confronto tra grafi PESATI

per la somiglianza valgono come sempre criteri soggettivi (da qui l'importanza del relative feedback): dò più importanza alla durata totale oppure alla durata relativa tra le singole parti oppure ancora alla somiglianza tra i singoli oggetti?

--> è importante il feedback dell'utente per guidare il sistema verso l'interpretazione giusta!

How to compare OCPNs



Maria Luisa Sapino (BDMM 2010)

27
7

===== 226 =====

TEXTURE: forme che si ripetono in modo più o meno regolare, eventualmente associate a delle direzionalità; alcune hanno direzionalità dominante, altre no ma hanno forme dominanti, ecc... danno l'idea del tipo di superficie (es: cemento, erba, terreno)

TESSITURA: una delle feature delle immagini che ci interessa catturare, contribuisce a definire la somiglianza tra immagini

per rappresentare la tessitura passiamo alla rappresentazione dell'immagine in scala di grigi -> come la variazione di luminosità è distribuita sulla superficie

come si rappresenta?
* direzionalità
* linearità
* regolarità/omogeneità della superficie
* granularità (dimensione degli elementi che si ripetono)

- rappresentazione del COLORE: è sufficiente procedere pixel per pixel
- diverso per la TESSITURA: nessun pixel da solo mi dà info sulla tessitura, non è una proprietà locale di un pixel ma è proprietà globale di finestra di pixel -> variazione di intensità e direzionalità del segnale luminoso nell'immagine
(N.B. poiché è "variazione" è necessario confronto tra almeno 2 pixel!)

metodi diversi per rappresentare (a differenza dei colori non c'è uno standard) -> il modello da utilizzare dipende dall'applicazione o dal tipo di pattern che si cerca o che si vuole rappresentare:

- modelli statistici: immagine vista come dominio in cui statisticamente a ciascun punto dell'immagine si associa un valore che corrisponde al suo livello di grigio
- rappresentazioni frattali: funzioni che generano forme le cui sottoforme hanno aspetto simile a macroforma di partenza (es: albero e rami di un albero, rametti dei rami, ecc... - anche i fiocchi di neve)

in generale se non conosciamo regolarità di questo tipo si cerca di misurare le variazioni di intensità luminosa di pixel adiacenti e si rappresentano tramite istogramma delle tessiture le variazioni osservate

ESEMPIO:
abbiamo immagine in scala di grigi di cui vogliamo ottenere la tessitura
* per ogni pixel dell'immagine applico operazione di convoluzione con 2 specifiche matrici che servono a misurare differenze su asse X e su asse Y: per misurare la variazione rispetto all'asse X, confronto ciascun pixel con la situazione alla sua DX e alla sua SX, lasciando invariata la situazione sulla verticale:

$$DX = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix}$$

--> conto come variazione negativa quella verso il sx e positiva quella verso dx

- idem (e opposto) per l'asse Y:

$$DY = \begin{bmatrix} 1 & 2 & 1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{bmatrix}$$

--> conto come variazione negativa quella verso il basso e positiva quella verso l'alto

COSA FACCIO PER OGNI PUNTO CONSIDERATO?

* misuro l'esistenza di variazioni di luminosità con i pixel adiacenti per capire se sia un bordo: come?
- calcolo il prodotto della finestra 3x3 che ha al centro il pixel considerato per la matrice DX

se considero matrice di texture

$$\begin{bmatrix} 100 & 0 & 0 \\ 100 & 100 & 0 \\ 0 & 100 & 100 \end{bmatrix}$$

prima calcolo variazione rispetto a x -> $dx = 100 \times (-1) + 0 \times 0 + 1 \times 0 + 100 \times -2 + 100 \times 0 + 0 \times 2 + 0 \times -1 + 100 \times 0 + 100 \times 1 = -100 -200 + 100 = -200$

faccio la stessa cosa per la variazione rispetto a y -> $dy = -200$

* **INTENSITA'**: posso calcolare l'intensità semplicemente combinando le variazioni lungo i due assi: $\sqrt{dx^2 + dy^2}$ -> però la sola intensità non è informativa, mi dà solo il modulo del vettore di variazione ma mi manca la direzione!

* **DIREZIONE**: 'calcolando $\arctan(dy / dx)$ ottengo l'angolo secondo la quale la variazione si manifesta

--> ad ogni pixel si associa un GRADIENTE, cioè un vettore descritto da una coppia [intensità, direzione] di variazione dell'intensità luminosa

bisogna ancora fare qualche operazione di pulizia:

1) dopo aver calcolato la variazione dell'intensità di luminosità, si filtra togliendo i pixel di "disturbo" che sono sotto una certa soglia di luminosità (percentuale o threshold - es: sotto al 5% dell'intensità globale) oppure che sono "situazioni isolate"

2) lavoro di assottigliamento dei bordi: se due pixel adiacenti hanno lo stesso comportamento allora se ne conta una solo

--> abbiamo modellato la texture mediante un insieme di TERNE (pixel P, intensità di variazione su P, direzione della variazione su P), o alternativamente come COPPIA (pixel P, gradiente di P) in cui il gradiente è rappresentato come una coppia

ORA POSSO FARE ISTOGRAMMA DELLE TESSITURE!

- asse per valore di intensità

- asse per valore di direzione

--> a ciascun punto corrisponde contatore dei pixel -> è un istogramma bidimensionale

volendo posso aggiungere una terza dimensione:

- la distanza dal più vicino omologo, ossia dal punto più vicino che presenta la stessa variazione -> permette di rappresentare la granularità della tessitura, perché dà anche informazione sulla frequenza

==== 228 ====

rappresentazione delle FORME: identificare regioni di immagine che siano omogenee come colore, identificare i confini di queste regioni e descrivere opportunamente la forma che questi segmenti rappresentano -> ovviamente è processo molto più difficile rispetto a descrivere colore e tessitura, perché introduce componente semantica

COME RICONOSCERE I BORDI?

1) segmentare l'immagine, ovvero identificare porzioni di immagini che al loro interno sono omogenee (hanno lo stesso colore)

* si può fare con algoritmi di CLUSTERING -> fare CLUSTERING: raggruppare in cluster (= gruppi omogenei) oggetti simili e che sono plausibilmente diversi da oggetti che sono in altri cluster (N.B. ovviamente bisogna prima definire la somiglianza tra oggetti)

--> posso clusterizzare mettendo nello stesso cluster i pixel adiacenti che hanno lo stesso colore

* altro modo: a partire da un pixel si fa crescere la regione attorno fino a che non si può più allargare perché il pixel che si incontra è diverso -> ho riconosciuto segmento di immagine

* altro modo ancora: WATERSHED, basato sullo studio del gradiente e della differenza di luminosità tra due pixel -> vedere l'immagine come una superficie topologica, non è superficie piana ma alcune parti sono più alte rispetto ad altre: è una superficie 3D -> il riconoscimento della forma viene modellato immaginando di riempire la superficie con acqua: dove l'acqua si addensa si delineano forme! la pendenza è data dalla direzione del gradiente

2) una volta rilevati i confini con una di queste tecniche bisogna rappresentarli: come confrontare le forme?

tecnica di CHAINING serve per rappresentare e quindi poi poter fare confronti:

* si definiscono 8 direzioni rispetto a cui ci si può muovere nella matrice (= 8 pixel attorno a quello su cui ci si sta concentrando)

* si sceglie una delle 8 direzioni

* si percorre in senso orario la curva indicando ogni volta quale è la direzione che si è dovuta seguire per passare da un pixel a quello successivo nella catena, e si aggiunge il codice associato alla direzione scelta, fino ad arrivare ad una curva chiusa -> così la curva si può codificare con una stringa di numeri interi, lunga quanto il numero di pixel che si trovano sulla curva

(continua su slides)

Ci sono tre tipi di testo:

- Testo libero: testo che non ha una struttura sintattica o semantica, ad esempio email, libri, ecc.

.Exact keyword query: trovare nel testo l'occorrenza di una parola

.Somiglianza sintattica: verifica le differenze a livello sintattico tra due testi, es. per verificare se due sequenze di DNA sono simili, es. con distanza di EDIT

.Somiglianza semantica: es. per libri è più interessante cercare differenze semantiche piuttosto che sintattiche. Ad esempio 'mamma' e 'madre' sono sintatticamente distinti ma hanno lo stesso significato

- Testo semistruzzurato: ha una sintassi semplice e non vincolante come in xml dove all'interno dei tag (regolati da una sintassi) si può inserire testo libero, quindi senza una particolare sintassi

- Testo strutturato: testo con una sintassi rigida

Come per le immagini è necessaria l'estrazione di un surrogato.

Vedremo solo query su testo non strutturato.

Text

- Unstructured text (free text)
 - Exact keyword query
 - Syntactically similar keyword query
 - Semantically similar keyword query
- Semistructured text (SGML,XML)
- Structured text
 - Structural similarity query

28
1

Maria Luisa Sapino (BDMM 2010)

Keyword based retrieval

- Given a keyword, find all documents that contain the keyword
- Inverted indices when word boundaries are known
 - Use B-trees for exact retrieval
 - Use "trie"s and suffix trees for prefix based retrieval
- Need substring search when word boundaries are not known
 - exact: Bayer, Moore (BM) or Knuth,Morris,Pratt (KMP)
 - syntactic similarity: Wu,Manber

28
2

Maria Luisa Sapino (BDMM 2010)

Documento testuale =(rappresentato come) multiinsieme di parole chiave (multiinsieme una parola può occorrere più volte). Ogni parola nel testo può essere

- Content word: porta informazione es. madre, famiglia, città
- Stop words: articoli, preposizioni, ausiliari -> non portano informazione e vengono tipicamente trascurate in fase di costruzione del surrogato

Per costruire un surrogato occorre eseguire una fase di Preprocessing:

- Si eliminano le Stop Words (non portano informazione)

- Stemming: riduce alla stessa rappresentazione di parole che sono varianti sintattiche tra loro, es. mamma, mamme -> mamm
- Lemmatizzazione: si tiene un rappresentante di una parola per una categoria di parola con una semantica simile ad es. andare, andai, vado -> andare oppure mamma,mamme -> mamma (rappresentante)

- Phrasing: riconoscimento di sequenze di parole composte come unica entità che ha un certo significato. Ad esempio ogni occorrenza di 'torri gemelle' viene considerata come parola a se stante e non aumenta le occorrenze di 'torri' ne 'gemelle' nel conteggio delle keyword

Text (as a collection of keywords).....

- Each document is represented as a multi-set of keywords
 - Content words (terms)
 - Non-content words (stop words)
- Preprocessing
 - Stop word removal: eliminates stop words
 - Stemming: identifies roots of the terms
 - Phrasing: identifies compound terms

28
3

Maria Luisa Sapino (BDMM 2010)

Zipfian Distribution

- The frequency of the k^{th} most frequent word in a collection is $(1/k)^{\theta}$ times the most frequent word.

Maria Luisa Sapino (BDMM 2010)

28
4

Zipfian Distribution

- The frequency of the k^{th} most frequent word in a collection is $(1/k)^{\theta}$ times the most frequent word.

$$\text{freq}(k_i) = 1/(r_i \ln(1.78V))$$

Maria Luisa Sapino (BDMM 2010)

28
5

Zipfian Distribution

- The frequency of the k^{th} most frequent word in a collection is $(1/k)^{\theta}$ times the most frequent word.

$$\text{freq}(k_i) = 1/(r_i \ln(1.78V))$$

Maria Luisa Sapino (BDMM 2010)

28
6

La frequenza di parole nel testo seguono la distribuzione di Zip.

Il grafico ordina le parole per frequenza sull'asse delle ascisse (rank). In generale nei testi la parola k -esima su tale asse ha frequenza $(1/k)^{\theta}$ rispetto alla parola più frequente.

Il grafico mostra che:

- Ci sono poche parole molto frequenti in un testo. Queste tipicamente le stop words che vengono dunque eliminate
- Ci sono molte parole poco frequenti in un testo, ma quelle che sono sotto una certa frequenza in un testo (calcolata con num occorrenze parola / parole totali nel testo) sono di poca importanza e vengono trascurate.
- Interessa dunque principalmente la coda della curva

r_i : ranking

V : dimensione del vocabolario

Esistono online dei database che contengono le stop word in una lingua, anche se bisogna tenere conto anche che in alcuni casi particolari le stop word possono essere significative, ad esempio tesi sull'importanza dell'articolo 'il' nella poesia romanesca: la stop word 'il' è molto importante. Quindi è importante adeguare l'uso di questi vocabolari al contesto.

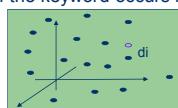
Vector representation

- Given a set of keywords, each document is represented as a vector:

$$d_i = \langle w_{i1}, w_{i2}, w_{i3}, \dots, w_{in} \rangle$$

where

- $w_{ij} = 0$, if the keyword does not occur in d_i
- $w_{ij} > 0$, if the keyword occurs in d_i



Maria Luisa Sapino (BDMM 2010)

28
7

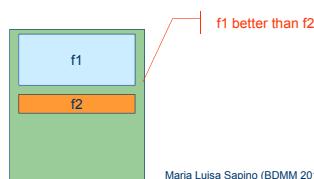
Testo =(rappresentato come) vettore in cui ogni dimensione è un termine (ad es. della lingua italiana), lemmatizzato o stemmatizzato (ad es. madre e madri avranno la stessa dimensione). Ogni dimensione è 0 se il termine corrispondente non è presente, > 0 altrimenti.

Siccome nella realtà un vettore di questo tipo contiene tanti zeri (un testo non contiene tutte le parole di una lingua), questi vettori $\langle w_1, \dots, w_n \rangle$ si rappresentano come liste di coppie ordinate $(indice, w_indice)$ così se si cerca la parola con indice i , si verifica che nella lista ~~sta presente la coppia (i, w_i) , se lo è il peso della parola i è w_i , altrimenti il peso è zero (quindi si risparmia spazio).~~

Occorre associare un peso ad ogni parola in base alla rilevanza che ha tale parola nel testo.

What are the weights????

- They need to capture how
 - good the term (feature) is in describing the content of the object



Maria Luisa Sapino (BDMM 2010)

28
8

tf: term frequency

n: numero di occorrenze del termine nel testo

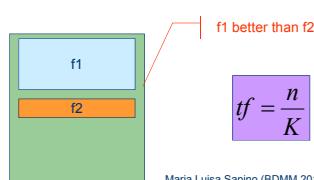
K: dimensione del testo

Come associare un peso ad un termine?

Occorre considerare due fattori

What are the weights????

- They need to capture how
 - good the term (feature) is in describing the content of the object



Maria Luisa Sapino (BDMM 2010)

28
9

- Usare la tf: parole che ~~occorrono molte volte~~ possono determinare bene l'argomento di cui parla il testo e quindi ~~hanno una grande importanza~~.

What are the weights????

- They need to capture how
 - differentiating the term (feature) is..



29
0

Maria Luisa Sapino (BDMM 2010)

2) Inverse document frequency: si basa sul fatto che parole che occorrono in pochi documenti del database sono più discriminanti, ossia portano maggiore entropia.

Ad esempio f2 permette di ottenere 3 documenti su 7 (se si cercano documenti con f2) mentre la ricerca su f1 restituisce l'intero database, per cui f2 è più importante di f1.

What are the weights????

- They need to capture how
 - differentiating the term (feature) is..



$$idf = \log\left(\frac{N}{m}\right)$$

Maria Luisa Sapino (BDMM 2010)

N: numero totale di documenti

m: numero di documenti in cui la feature occorre

idf: inverse document frequency

What are the weights????

- They need to capture how
 - good the term (feature) is in describing the content of the object
 - differentiating the term (feature) is..

$$tfidf = \frac{n}{K} \log\left(\frac{N}{m}\right)$$

29
2

Maria Luisa Sapino (BDMM 2010)

E' importante considerare entrambi gli aspetti, 1 e 2, quindi se ad esempio un termine è poco frequente in un documento ma è poco frequente anche nel DB allora l'importanza è più elevata. Quindi frequenza e IDF vanno combinate (considerati sempre assieme per stabilire l'importanza di un termine).

Il risultato è il calcolo di tfidf.

$$tfidf = \text{frequenza} * idf$$

Questa grandezza è globale, ossia è di carattere generale ed è dipendente dalla dimensione del DB (più il database è grande più è grande la dimensione).

Occorre normalizzare.

What are the weights????

- They need to capture how
 - good the term (feature) is in describing the content of the object
 - differentiating the term (feature) is..

$$\text{norm_tfidf} = \frac{n}{K} \frac{\log(\frac{N}{m})}{\max idf}$$

Idf of the keyword

Maria Luisa Sapino (BDMM 2010)

29
3

maxidf = massimo valore di idf riscontrato nel database
Quindi il secondo rapporto nella formula in questa slide ha range 0-1.

Il range di norm_tfidf è [0,1].

La normalizzazione è importante per avere sempre la stessa scala di misurazione della rilevanza, cosa importante soprattutto quando si ha a che fare con confronti con elementi di database diversi. Ad esempio se in DB1 il range dei pesi è 0-0.7 e in DB2 il range è 0-50, non si può dire in assoluto quando una feature sia rilevante, ma solo quanto lo sia rispetto ad altro all'interno dello stesso DB. Quindi non si possono confrontare feature di DB1 e di DB2.

Experiment results suggest that

- Poor terms have high document frequency
- Good terms have low document frequency
 - Problem: may not be queried often enough to be useful
- Best terms have medium document frequency

Maria Luisa Sapino (BDMM 2010)

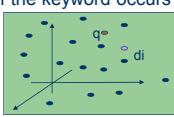
29
4

E' importante notare che anche se le parole meno frequenti in un database hanno molta importanza, alcune di esse sono anche cercate rarissimamente per cui in questi casi sono quasi inutili.

Occorre trovare un tradeoff tra probabilità di ricerca della parola e rilevanza nel database/documento.

How about query terms??

- Given a set of keywords, each query is also represented as a vector:
 $q = \langle w_{q1}, w_{q2}, w_{q3}, \dots, w_{qn} \rangle$
where
 - $w_{qj} = 0$, if the keyword does not occur in d_q
 - $w_{qj} > 0$, if the keyword occurs in d_q



Maria Luisa Sapino (BDMM 2010)

29
5

Ci sono due tipi di query su testo.

- Per chiave: ricerca di testi che contengono determinate parole

- Per somiglianza: ricerca testi simili ad un determinato testo. Molto simile a query di somiglianza su immagini: anche in questo caso si utilizzano i surrogati per verificare la somiglianza. Come per le immagini: dato in input il testo della query (es. per query trova tutti i testi simili a T1.txt il testo della query è T1.txt), si costruisce il surrogato e si effettuano i confronti tra questo surrogato e quelli presenti nel database.

How about query terms??

- They need to capture how
 - good the term (feature) is in describing the query
 - differentiating the term (feature) is...
 - Salton&Buckley suggests..

$$tfidf = \left(0.5 + 0.5 \frac{\frac{n}{K}}{\max freq} \right) \log\left(\frac{N}{m}\right)$$

Maximum term frequency in the query Frequency of the term in the query

Maria Luisa Sapino (BDMM 2010)

29
6

Are keywords independent??

29
7

Maria Luisa Sapino (BDMM 2010)

Are keywords independent??

- Vector model assumes that they are..
- ..but are they really?

29
8

Maria Luisa Sapino (BDMM 2010)

Tuttavia spesso si preferisce privilegiare il fatto che è presente un elevato numero di parole della query in un documento (anche se con bassa frequenza) piuttosto che sono presenti poche parole ma con frequenza elevata.

Ad esempio se la query comprende le parole 'partita' e 'calcio', è meglio considerare più simile un documento che contiene due occorrenze di 'partita' e una di 'calcio' piuttosto che uno che contiene 10 occorrenze di 'partita' e nessuna di 'calcio' (infatti in tal caso 'partita' potrebbe essere associata a 'basket').

Quindi i pesi del vettore della query (TFIDF) hanno valore minimo 0.5 (in questo modo testi che non contengono elementi della query hanno già una distanza di 0.5 per tali componenti) e l'elemento più frequente del testo della query ha valore 1 (max freq è il massimo valore di frequenza di un termine nella query).

NB: se si utilizza la formula in cui si normalizza la frequenza (rapporto tra il logaritmo e max idf) anche qui occorre normalizzare, altrimenti si farebbero confronti in misure fatte su scala differente.

La distanza coseno è la misura migliore della distanza per il testo perché privilegia testi che contengono le stesse parole della query. Infatti se un testo ha gli stessi termini (o quasi) del testo della query anche con frequenze abbastanza diverse, tendono comunque ad avere un angolo abbastanza basso e quindi una certa somiglianza.

Tutto ciò che è stato detto riguardo ai confronti tra testi assume l'indipendenza tra chiavi, ma alcune parole chiave sono dipendenti tra loro e occorre gestire tali dipendenze.

Ad esempio gatto, gatti provengono dalla stessa radice, ossia 'gatt'. Quindi considerarle due parole distinte potrebbe portare a rilevare come differenti testi simili.

Are keywords independent??

- Syntactic similarity
 - Prefix relationship
 - "cat" vs. "catle"
 - Edit distance:
 - "table" vs. "cable": 1 (replace "t" with "c")
 - "table" vs. "bale": 2 (delete "t"; swap "a" and "b")

29
9

Maria Luisa Sapino (BDMM 2010)

- Somiglianza sintattica: due parole sono sintatticamente simili, ad esempio se misurate con la distanza di edit.

Semantic relationships

- Corpora-independent
 - Synonym
 - Different but same meaning
 - Polysemy
 - More than one meaning
 - Hyponymy
 - K1 is an hyponym of K2 iff K1 is a K2
 - Hypernymy
 - K1 is an hypernym of K2 iff K2 is a K1
- Corpora-dependent
 - cooccurrence

30
0

Maria Luisa Sapino (BDMM 2010)

- Distanze semantiche: sono più importanti.
Si suddividono in due tipi
. Distanze indipendenti dal contesto (cioè dai documenti su cui si lavora):
-> Sinonimia, polesemia, iponimia (termine più specifico di un altro), iperonimia (termine più generale di un altro)

Occorre tenere conto di queste dipendenze e quindi ~~parole con queste dipendenze possono finire ad esempio in uno stesso cluster~~. Ad esempio si possono considerare le parole casa, villa, dimora, residenza come uguali così ad esempio la ricerca 'casa in vendita' restituisce testi che contengono 'casa', 'dimora' ecc.

. Distanze dipendenti dal contesto: relazione tra due parole causata dal fatto che queste occorrono entrambe (co-occorrono) in molti documenti

Semantic distance

- How dissimilar two terms are?
 - $\text{dist}(\text{man}, \text{woman})?$
 - $\text{dist}(\text{man}, \text{child})?$
 - $\text{dist}(\text{man}, \text{human})?$

30
1

Maria Luisa Sapino (BDMM 2010)

Semantic distance

30
2

Maria Luisa Sapino (BDMM 2010)

Per calcolare la distanza semantica tra due parole occorre avere una tassonomia semantica delle parole, allo scendere di livello dell'albero le parole sono più specifiche ed i figli di ogni nodo sono parole semanticamente più specifiche del loro padre (es. mammifero può avere come figli cane, uomo, ecc.)

- 1) La distanza può essere calcolata come il numero di archi da attraversare per passare da un nodo all'altro (dove ogni nodo è una parola)

es. $\text{dist}(\text{uomo}, \text{donna}) = 2$
 $\text{dist}(\text{uomo}, \text{animale}) = 3$

Questa distanza non tiene conto ne della frequenza di termini nei documenti ne della co-occorrenza di termini

esempio: se cane e gatto si trovano spesso assieme nei documenti mentre cane e scimmia no, la distanza tra cane e gatto dovrebbe essere minore di quella tra cane e scimmia mentre questo aspetto non è catturato da questo metodo di misurazione della distanza

Semantic distance

30
3

Maria Luisa Sapino (BDMM 2010)

- 2) Si calcola il contenuto informativo di ciascun nodo.

$$\text{Contenuto_informativo} = \log(1/f)$$

La distanza tra due oggetti è data dal contenuto informativo massimo degli antenati in comune.

In questo modo si tiene conto delle frequenze delle parole nei vari documenti.

PROBLEMA: dato un nodo v , la distanza tra due suoi figli u_1, u_2 è la stessa della distanza tra u_1 e qualsiasi elemento del sottoalbero di u_2 (discendente di u_2) - perché i loro antenati in comune sono gli stessi. Ad esempio pesce e mammifero hanno la stessa distanza di pesce e uomo.

Semantic distance (P. Resnick)

30
4

Maria Luisa Sapino (BDMM 2010)

Semantic distance (P. Resnick)

Class hierarchy

$\text{sim}(\text{fish}, \text{man}) = \max[\inf_{\text{content}}(\text{ca}(\text{man}, \text{fish}))]$

Note: $\text{sim}(\text{fish}, \text{man}) = \text{sim}(\text{fish}, \text{bird})$

30
5

Semantic distance (Richardson et al.)

```

graph TD
    everything --> alive
    everything --> notalive[not-alive]
    alive --> animal
    alive --> plant
    alive --> fungi
    alive --> ...
    ...
    animal --> bird
    animal --> fish
    animal --> mammal
    plant --> ...
    ...
    fungi --> ...
    ...
    mammal --> dog
    mammal --> cat
    mammal --> human
    dog --> man
    dog --> woman
    cat --> man
    cat --> woman
    human --> ...
    ...
  
```

Weight each edge

- density of the hierarchy
- depth of the edge
- information content of each end

Per risolvere questo problema
3) Si associa un peso ad ogni arco che dipende da tre fattori

- densità della gerarchia (numero di nodi del sottoalbero individuato dall'arco)
 - profondità dell'arco
 - contenuto informativo alle estremità dell'arco

Semantic distance (Richardson et al.)

Weight each edge

- density of the hierarchy
- depth of the edge
- information content of each end

$\Delta a < \Delta b$ (density)
 $\Delta a < \Delta c$ (depth)
 $\Delta a < \Delta d$ (ic)

- Delta a < Delta b perché la densità rilevata sull'arco a è maggiore di quella rilevata sull'arco b e quindi a maggior densità corrisponde un maggiore contenuto informativo (un oggetto che ha una tassonomia ~~densa descrive molto caratteristiche del concetto che rappresenta, descrive molto~~)
- Delta a < Delta c perché la profondità di a è maggiore e quindi i nodi hanno un contenuto informativo maggiore
- Delta a < Delta d perché è minore la differenza dei loro contenuti informativi

Term-to-term correlation

- Computes relationships between keywords given a corpora of documents
- Keyword connection matrix

30
8

Maria Luisa Sapino (BDMM 2010)

Criterio per misurare la co-occorrenza tra due termini in un documento

i,l: due keyword

c_{i,l}: stima del grado di co-occorrenza dei termini nei documenti del database

n_{i,l}: probabilità che le due keyword occorrono entrambe in un documento

n_i: probabilità che la keyword i occorra in un documento

n_j: come n_i ma per la keyword j

n_{i+l} - n_{i,l}: probabilità che solo una delle due parole chiave occorrono in un documento

La matrice di connessione delle keyword contiene dunque c_{i,l} in ogni cella (i,j)

Vector model

- Given a set of keywords, each document is represented as a vector:
 $d_i = \langle w_{i1}, w_{i2}, w_{i3}, \dots, w_{in} \rangle$
- We already discussed the salient features of this model...

30
9

Maria Luisa Sapino (BDMM 2010)

Extended Boolean Model

- Salton, Fox, Wu(83)

31
0

k1 or k2 = Δ

k1 and k2 = $\sqrt{2} - \Delta$

Maria Luisa Sapino (BDMM 2010)

Extended Boolean Model: serve a misurare la generalità e specificità di documenti testuali. Supponiamo di avere un vocabolario di due termini k_1 e k_2 . Il grafico a sinistra nel lucido ha come ascissa il peso del termine k_1 e come ordinata il peso di k_2 , quindi denota lo spazio documenti con due termini (k_1 e k_2). Se un documento è rappresentato dal vettore $(0,0)$ significa che i termini k_1 e k_2 non occorrono in tale documento, per cui non si menziona né k_1 , né k_2 ossia $\neg k_1 \text{ and } \neg k_2 = (\text{De Morgan}) \neg (k_1 \text{ or } k_2)$.

Maggiore è la lunghezza Delta del vettore che collega l'origine al punto che rappresenta il documento, più il documento è distante da not (k_1 or k_2), e quindi è sempre più vicino a k_1 or k_2 , ossia si parla di k_1 o k_2 . Il grado di orness (quanto i due termini sono in disgiunzione) rappresenta la generalità del documento: es. per i termini cane e gatto, se il grado di orness per cane or gatto è elevato, significa che il documento esprime un concetto o un altro, per cui è generale ossia è vicino all'elemento padre della tassonomia (animale) mentre al contrario il grado di andness cane and gatto esprime specificità.

Il grafico a destra mostra il duale, ossia il grado di andness. Tanto maggiore è Delta quando il documento è meno specifico.

Probabilistic Model

- Robertson&Jones(76)
 - Binary Independence Model
- Given a query and a document, estimate the probability that the user will find the document interesting
 - Assumption: there is an ideal set!! Can we estimate the properties of the ideal set.
 - We will come back to this model later..

Relevance feedback!!!

Maria Luisa Sapino (BDMM 2010)

31
1

Modello probabilistico: si utilizza per valutare la probabilità che l'utente sia interessato ad un risultato rispetto alla query. La difficoltà sta nel fatto che non si conoscono a priori in modo esatto le preferenze dell'utente per cui non è possibile calcolare tra le probabilità esattamente.

L'utente dovrà dunque fornire informazioni per raffinare il calcolo delle probabilità.

Fuzzy Set Model

- Each query term defines a fuzzy set
- Each document has a degree of membership in this set
- Example: membership degree of document d_j in keyword k_i

$$\mu_{k_i, j} = 1 - \prod_{k_l \in d_j} (1 - c_{i, l})$$

- We will come back to this model later

Query processing!!!

Maria Luisa Sapino (BDMM 2010)

31
2

Lo vedremo bene quando parleremo delle query.

Ogni query ha un insieme di risultati (fuzzy set) e ogni oggetto (testo, immagine, ecc.) ha un grado di appartenenza a tale insieme: 0 non appartiene, 0.5 appartiene abbastanza, 0.6 appartiene ancora di più, ..., 1 appartiene con certezza all'insieme.

Formula per calcolare la correlazione tra una chiave k_i e un documento d_j :

1 - $c_{i,j}$: grado di co-occorrenza delle chiavi i e j

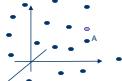
2 - $c_{i,j}$: grado di non correlazione tra le chiavi i e j

3 - produttoria: misura quanto la chiave è indipendente dalle altre nel documento (unione di tutti i casi in cui la chiave i è indipendente da una nel documento).

4 - (1 - produttoria): grado di correlazione delle chiavi con il documento, quindi grado di appartenenza della chiave i al documento j .

Challenge....

- Traditional data is **one dimensional**.
- Multimedia data is **multi dimensional**.
 - Ex. Maps are 2D
 - In general, if a given information has k features, it can be represented by a k-dimensional space



Maria Luisa Sapino (BDMM 2010)

43
4

What kind of queries we can expect?

- Given a set of point in k-dimensional space
 - Exact match:
 - find if a given point is in the set or not
 - Nearest neighbor:
 - find the closest point to a given point
 - Range search:
 - Given a region (rectangle or circle), find all the points in the given region

Maria Luisa Sapino (BDMM 2010)

43
5

General approach

- Divide the space into regions
- Insert the new object into the corresponding region
- If the region is full, split the region
- retrieval: determine which regions are required to answer a given query and limit the search to these regions

Maria Luisa Sapino (BDMM 2010)

43
6

Obiettivo degli indici: trovare gli oggetti cercati dalla query senza dover ricercare in tutto il database.

L'indice (tipicamente è un albero) deve permettere di potare sottoalberi in modo tale da permettere, durante la ricerca, di localizzare il sottoalbero che contiene tutti i risultati migliori della query.

Esistono fondamentalmente tre tecniche per risolvere questo problema:

- Clustering: raggruppamento di oggetti simili in un unico cluster.
Viene utilizzato per le query di somiglianza.

- Classificazione: associa oggetti a proprietà

- Indicizzazione: stabilisce un ordine tra i dati in modo tale da localizzare velocemente i dati di interesse. Tipicamente gli indici sono alberi.

Come fare gli indici su dati multidimensionali (che sono a più dimensioni)?

1) fare un indice per ogni feature. Problema -> troppo costoso e si vogliono confrontare gli oggetti per più dimensioni e non per una alla volta. Si vuole un indice che permetta di ordinare gli oggetti con tutte le dimensioni.

Tipi di operazioni in cui si utilizzano queste strutture di supporto (indici, cluster, ecc.)

1) Inserimento

2) Cancellazione

3) Query

a) trovare oggetto dato nel DB

b) k-NN o top-k: k oggetti più simili a quello dato

c) query di range: oggetti che distano al più \delta dalla query

Per costruire gli indici dobbiamo suddividere lo spazio del database in regioni. Di fatto gli indici a una dimensione (come quelli sul modello relazionale, che indicizzano per un attributo, quindi una feature e non più feature) suddividono lo spazio monodimensionale dalla radice in n segmenti e poi ancora in sottosegmenti, ecc.

Per indici multidimensionali si suddivide lo spazio in regioni.

Ogni punto è inserito in una posizione dello spazio che dipende dai valori delle sue feature. La regione che contiene quel punto è la regione di appartenenza del punto.

Ogni regione può contenere un numero limitato di oggetti. Quando la regione si riempie, la si spezza in due regioni..

RETRIEVAL: si determinano le regioni necessarie per rispondere alla query (potenzialmente sono una risposta alla query). Solo quelle devono essere analizzate.

Is there an alternative to multi-dimensional space decomposition?

- YES!
 - Convert a given k-D space to 1D space
 - We know how to handle 1D space!!
- Don't we loose information??
 - Yes, but if we are careful, we can minimize the information loss.

43
7

Maria Luisa Sapino (BDMM 2010)

Ma non si può mappare gli oggetti in k-dimensioni in uno spazio monodimensionale così da poter utilizzare le strutture note (B-Tree, B+ Tree, ecc.)?
Si perdono delle informazioni, però posso minimizzare tale perdita?

Space filling curves

- Convert a k-D space into 1D space such that points that are close to each other in k-D space are also close to each other in 1-D space

43
8

Maria Luisa Sapino (BDMM 2010)

Possibile soluzione: SFC

partire da spazio di dimensione k e convertirlo in uno spazio di dimensione 1 definendo un ordine totale tra tutti i punti nello spazio a k-dimensione (si immaginino delle curve che attraversano in quell'ordine tutti i punti)

Row order/column order

0	1	2	3	4	5	6	7
0	1	2	3	4	5	6	7
8	9	10	11	12	13	14	15
8	9	10	11	12	13	14	15
16	17	18	19	20	21	22	23
16	17	18	19	20	21	22	23
24	25	26	27	28	29	30	31
24	25	26	27	28	29	30	31
32	33	34	35	36	37	38	39
32	33	34	35	36	37	38	39
40	41	42	43	44	45	46	47
40	41	42	43	44	45	46	47
48	49	50	51	52	53	54	55
48	49	50	51	52	53	54	55
56	57	58	59	60	61	62	63
56	57	58	59	60	61	62	63
64	65	66	67	68	69	70	71
64	65	66	67	68	69	70	71
72	73	74	75	76	77	78	79
72	73	74	75	76	77	78	79
80	81	82	83	84	85	86	87
80	81	82	83	84	85	86	87
88	89	90	91	92	93	94	95
88	89	90	91	92	93	94	95
96	97	98	99	100	101	102	103
96	97	98	99	100	101	102	103
104	105	106	107	108	109	110	111
104	105	106	107	108	109	110	111
112	113	114	115	116	117	118	119
112	113	114	115	116	117	118	119
120	121	122	123	124	125	126	127
120	121	122	123	124	125	126	127
128	129	130	131	132	133	134	135
128	129	130	131	132	133	134	135
136	137	138	139	140	141	142	143
136	137	138	139	140	141	142	143
144	145	146	147	148	149	150	151
144	145	146	147	148	149	150	151
152	153	154	155	156	157	158	159
152	153	154	155	156	157	158	159
160	161	162	163	164	165	166	167
160	161	162	163	164	165	166	167
168	169	170	171	172	173	174	175
168	169	170	171	172	173	174	175
176	177	178	179	180	181	182	183
176	177	178	179	180	181	182	183
184	185	186	187	188	189	190	191
184	185	186	187	188	189	190	191
192	193	194	195	196	197	198	199
192	193	194	195	196	197	198	199
200	201	202	203	204	205	206	207
200	201	202	203	204	205	206	207
208	209	210	211	212	213	214	215
208	209	210	211	212	213	214	215
216	217	218	219	220	221	222	223
216	217	218	219	220	221	222	223
224	225	226	227	228	229	230	231
224	225	226	227	228	229	230	231
232	233	234	235	236	237	238	239
232	233	234	235	236	237	238	239
240	241	242	243	244	245	246	247
240	241	242	243	244	245	246	247
248	249	250	251	252	253	254	255
248	249	250	251	252	253	254	255
256	257	258	259	260	261	262	263
256	257	258	259	260	261	262	263
264	265	266	267	268	269	270	271
264	265	266	267	268	269	270	271
272	273	274	275	276	277	278	279
272	273	274	275	276	277	278	279
280	281	282	283	284	285	286	287
280	281	282	283	284	285	286	287
288	289	290	291	292	293	294	295
288	289	290	291	292	293	294	295
296	297	298	299	300	301	302	303
296	297	298	299	300	301	302	303
304	305	306	307	308	309	310	311
304	305	306	307	308	309	310	311
312	313	314	315	316	317	318	319
312	313	314	315	316	317	318	319
320	321	322	323	324	325	326	327
320	321	322	323	324	325	326	327
328	329	330	331	332	333	334	335
328	329	330	331	332	333	334	335
336	337	338	339	340	341	342	343
336	337	338	339	340	341	342	343
344	345	346	347	348	349	350	351
344	345	346	347	348	349	350	351
352	353	354	355	356	357	358	359
352	353	354	355	356	357	358	359
360	361	362	363	364	365	366	367
360	361	362	363	364	365	366	367
368	369	370	371	372	373	374	375
368	369	370	371	372	373	374	375
376	377	378	379	380	381	382	383
376	377	378	379	380	381	382	383
384	385	386	387	388	389	390	391
384	385	386	387	388	389	390	391
392	393	394	395	396	397	398	399
392	393	394	395	396	397	398	399
400	401	402	403	404	405	406	407
400	401	402	403	404	405	406	407
408	409	410	411	412	413	414	415
408	409	410	411	412	413	414	415
416	417	418	419	420	421	422	423
416	417	418	419	420	421	422	423
424	425	426	427	428	429	430	431
424	425	426	427	428	429	430	431
432	433	434	435	436	437	438	439
432	433	434	435	436	437	438	439
440	441	442	443	444	445	446	447
440	441	442	443	444	445	446	447
448	449	450	451	452	453	454	455
448	449	450	451	452	453	454	455
456	457	458	459	460	461	462	463
456	457	458	459	460	461	462	463
464	465	466	467	468	469	470	471
464	465	466	467	468	469	470	471
472	473	474	475	476	477	478	479
472	473	474	475	476	477	478	479
480	481	482	483	484	485	486	487
480	481	482	483	484	485	486	487
488	489	490	491	492	493	494	495
488	489	490	491	492	493	494	495
496	497	498	499	500	501	502	503
496	497	498	499	500	501	502	503
504	505	506	507	508	509	510	511
504	505	506	507	508	509	510	511
512	513	514	515	516	517	518	519
512	513	514	515	516	517	518	519
520	521	522	523	524	525	526	527
520	521	522	523	524	525	526	527
528	529	530	531	532	533	534	535
528	529	530	531	532	533	534	535
536	537	538	539	540	541	542	543
536	537	538	539	540	541	542	543
544	545	546	547	548	549	550	551
544	545	546	547	548	549	550	551
552	553	554	555	556	557	558	559
552	553	554	555	556	557	558	559
560	561	562	563	564	565	566	567
560	561	562	563	564	565	566	567
568	569	570	571	572	573	574	575
568	569	570	571	572	573	574	575
576	577	578	579	580	581	582	583
576	577	578	579	580	581	582	583
584	585	586	587	588	589	590	591
584	585	586	587	588	589	590	591
592	593	594	595	596	597	598	599
592	593	594	595	596	597	598	599
596	597	598	599	600	601	602	603
596	597	598	599	600	601	602	603
604	6						

Row order/column order

44
0

Maria Luisa Sapino (BDMM 2010)

su spazio monodimensionale posso poi usare struttura già conosciuta, come B-tree

Row order/column order

44
1

Problems:
0-8
7-8

Maria Luisa Sapino (BDMM 2010)

(0, 8) risultano lontani -> causano MISSES
(7, 8) risultano vicini -> causano FALSE HITS

Row prime order/column prime order

44
2

Problems:
0-15

Not a problem:
7-8

Maria Luisa Sapino (BDMM 2010)

posso ridurre il # di casi critici ma non riesco ad eliminarli del tutto -> concentro la criticità in alcuni punti

Cantor diagonal order

Maria Luisa Sapino (BDMM 2010)

44
3

Diagonali di Cantor: rende abbastanza vicini nello spazio monodimensionale oggetti vicini nelle diagonali.

Z-order curve (hilbert curve)

Maria Luisa Sapino (BDMM 2010)

44
4

Rappresentazione a frattale della Z.

Questa tecnica più diffusa perché è semplicissimo il calcolo (da 2 dim. a 1).

Z-order curve (hilbert curve)

Easy to compute (bit-shuffling): $1(001) \times 2 (010) = 6 (000110)$

Maria Luisa Sapino (BDMM 2010)

44
5

Infatti basta eseguire un'operazione di bit shuffling:
dato il punto (a,b) con rappresentazioni in binario

$$a = a_1 \dots a_n$$

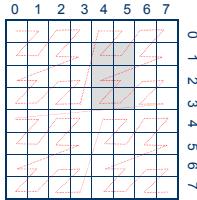
$$b = b_1 \dots b_m$$

$$\text{punto} = a_1 b_1 \dots a_n b_m \dots$$

Oltre ad essere semplice, questa misura è anche quella

che preserva maggiormente le distanze.

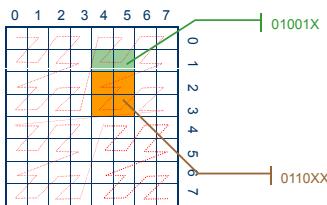
Z-order curve (hilbert curve)



44
6

Maria Luisa Sapino (BDMM 2010)

Z-order curve (hilbert curve)

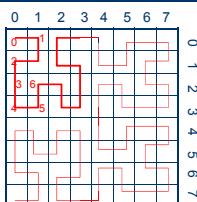


44
7

Range search can be implemented using tries...

Maria Luisa Sapino (BDMM 2010)

Peano-hilbert curve



44
8

Maria Luisa Sapino (BDMM 2010)

E' la misura migliore però è più difficile da calcolare.

In ogni caso le tecniche per convertire uno spazio a k-dimensioni in uno a una dimensione non sono ampiamente utilizzate perché si perdono informazioni durante la trasformazione (soprattutto sulle distanze tra punti).

Indexing

- What are we indexing???
 - Text → tries
 - Numbers, text → B-trees, B+ trees, B*trees
 - Images → ????????????
- Which feature are we going to index on?
 - Color? Texture? Time? (image series)
- What do we need to specify?
 - Lines? Points? Space?

45
0

Maria Luisa Sapino (BDMM 2010)

How do we index points?

- Given
 - a space of N-dimensions
 - M points
 - a distance function between points
- we can use multidimensional index structures
 - k-d trees
 - point quadtrees
 - MX quadtrees
 - R-trees
 - TV-trees
 - X-trees

45
1

Maria Luisa Sapino (BDMM 2010)

Un indice si costruisce a partire da

- spazio a N-dimensioni, con $N > 1$
- M punti (oggetti del database)
- funzione metrica di distanza

So...

- we can answer queries of the form
 - Given
 - a point X in N-dimensional space
 - Find
 - all points Y that are in its proximity ($d(X,Y) < \varepsilon$)

45
2

Maria Luisa Sapino (BDMM 2010)

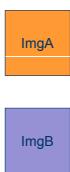
...thus...

- If
 - we represent any feature as a point in N-dimensional space (color, texture, shape, etc.)
 - we define a distance function between those points
 - (larger distance → lower similarity)
- Then
 - we can find media object with similar properties.

45
3

Maria Luisa Sapino (BDMM 2010)

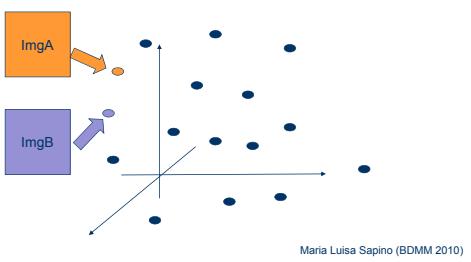
Populate database



45
4

Maria Luisa Sapino (BDMM 2010)

Populate database



45
5

Maria Luisa Sapino (BDMM 2010)

Map query image

Range search

The diagram shows a 2D coordinate system with x and y axes. A green circle, representing the search range for a query, is centered on a red dot labeled "A match". Several blue dots are scattered in the space, some within the green circle and some outside. Two boxes, "ImgA" (orange) and "ImgB" (purple), each have an arrow pointing to one of the blue dots located inside the green circle. The text "Range search" is displayed prominently at the top.

come organizziamo i dati in modo tale da consentire di rispondere alla query evitando (dove possibile) la ricerca esaustiva? -> vogliamo organizzazione che metta vicini (risp. lontani) oggetti vicini nello spazio (risp. lontani) e sfruttare la vicinanza di memorizzazione per fare pruning e scartare parti di db da analizzare

Grid File

lo spazio viene ripartito senza tenere conto della effettiva distribuzione dei dati

tecnica per organizzare informazione:
GRID FILES sono file in cui si distribuiscono
opportunamente gli oggetti del db rispetto ad una griglia che viene imposta nello spazio su cui gli oggetti sono distribuiti:
si memorizzano nella stessa pagina su disco i dati che cadono nella stessa cella -> il fatto di cadere nella stessa griglia cattura info sulla distanza reciproca tra due celle -> permette di conoscere distanza minima tra due celle non adiacenti

- + è struttura facile da definire
- rischio di sovradimensionare o sottodimensionare la cella
 - rischio di avere celle vuote e sprecare spazio
 - le query possono richiedere di toccare molte celle

Grid File

- Every cell is one disk page

Maria Luisa Sapino (BDMM 2010)

45
9

Point Trees

Point Trees: classi di alberi per indicizzare punti (vettori) multidimensionali.

L'idea di base è quella di suddividere lo spazio in regioni e di esplorarle arrivando alla regione di interesse.

How can we divide space?

- Let us assume that the space is 2-d
- There are many ways to divide the space
 - Fixed size squares
 - Triangles
 - Rectangles
 - Arbitrary space decomposition

• Each line divides the space into two

- Line: $n_1x + n_2y = c$
- Regions: $n_1x + n_2y \geq c$
 $n_1x + n_2y < c$

Maria Luisa Sapino (BDMM 2010)

46
1

Come suddividere le regioni? Deve essere un'operazione poco costosa.

es: suddivisione in quadrati di dimensione fissa, o triangoli, rettangoli, gridfile, suddivisione arbitraria, ecc.

Point quadtrees (Finkel and Bentley 74)

- Key features:
 - Every node in a point quadtree *implicitly* represents a rectangular region.
 - Each node contains an *explicit* point labeling it.
 - Root represents the whole region.
 - Each node's region is split into 4 parts ("quadrants") by drawing a vertical and a horizontal line through the point labeling the node.
 - Each node has 4 children corresponding to the 4 "quadrants" above.

46
2

Maria Luisa Sapino (BDMM 2010)

Point quadtree: point tree in cui nodo -> identifica una regione di spazio rettangolare (quindi è facile da individuare e rappresentare) tramite un punto ed è etichettato dalle sue coordinate.

Nodo (i,j) indica una regione suddivisa in 4 regioni delimitate da una retta orizzontale e una verticale che passano per quel punto (per convenzione le sottoregioni si denotano con NW, NE, SE, SW).

Il nodo (i,j) ha 4 figli, ciascuno dei quali identifica una delle 4 sottoregioni di (i,j) e a sua volta le suddivide altre 4 sottoregioni e così via...

Point quadtrees: example

(15,10)
Represents whole region
0,0

46
3

Maria Luisa Sapino (BDMM 2010)

per inserire un nuovo nodo si considera il quadrante in cui viene posizionato rispetto a quanto c'è già nell'albero
 $\rightarrow (10,14)$ è a N-O della radice $(15,10)$

per capire dove si trova il punto faccio il confronto tra le singole coordinate e guardo il segno

Point quadtrees: example

(15,10)
NW
(10,14)
0,0

46
4

Maria Luisa Sapino (BDMM 2010)

Point quadtrees: example

46
5

Maria Luisa Sapino (BDMM 2010)

(18,5) è a S-E rispetto a (15,10)

Point quadtrees: example

46
6

Maria Luisa Sapino (BDMM 2010)

(2, 12) è a N-W rispetto alla radice, ma il posto è già occupato da (10, 14): confronto con questo nodo e inserisco in base al quadrante

Observation

- The structure of the tree depends on the insertion order!!!!
- Exercise: try to insert nodes in the following order
(18,5) (15,10), (2,12) (10,14)
and compare the resulting tree with the previous one.

46
7

Maria Luisa Sapino (BDMM 2010)

* la struttura dipende dall'ordine di inserimento dei punti

* si può ottenere un albero molto sbilanciato (= caso estremo: lista come nell'esercizio proposto)

(18, 5)
NW-----
(15, 10)-----
NW-----
(2, 12)-----
NE-----
(10, 14)---

Key Points

- Suppose a point quadtree has N nodes in it.
- Worst case height = N.
- Worst case insertion time = N.
- Other operations are:
 - Deletion: delete a point
 - Range query: find all points within a given region
 - NN query: find the nearest neighbor (or M nearest neighbors) of a given point.

46
8

Maria Luisa Sapino (BDMM 2010)

quadtree con N nodi

- * nel caso peggiore l'altezza dell'albero è N, quindi anche il costo per l'operazione di inserimento
- * cancellazione di un nodo interno causa riorganizzazione dell'albero -> è possibile che si debbano spostare ogni volta più nodi

Con k dimensioni ogni nodo permette di creare 2^k suddivisioni

Deletion

- Suppose T is the root of a point quadtree and you want to delete (x,y).
- Steps:
 - Find (x,y) by doing a search.
 - If it is a leaf node, then simply set the appropriate link field of its parent to nil (and return the node to available storage).
 - What if it is not a leaf ?

46
9

Maria Luisa Sapino (BDMM 2010)

passi per CANCELLARE un nodo:

- * prima bisogna identificarlo nell'albero!
- se è una foglia si cancella
- se non è una foglia, bisogna ristrutturare l'albero in modo tale da mantenere le relazioni tra gli elementi e quindi le proprietà del quadtree

La cancellazione è molto costosa perché riporta ad una grande ristrutturazione dell'albero.

Soluzione: si associa un tag al nodo cancellato mantenendolo in memoria (ma non è più considerato un oggetto del database, quindi viene ignorato). Quando si raggiunge un certo numero di nodi con il tag (quindi cancellati dal database) si ricrea l'intero indice.

In ogni caso quando si rimuove un nodo in un sottoalbero occorre scegliere un una nuova radice (del sottoalbero). Se questo nodo non è una foglia nello spostarla come nuova radice occorre considerare un'operazione di cancellazione di quel nodo nel sottoalbero e quindi si genera una serie di cancellazioni ricorsive -> troppo costoso

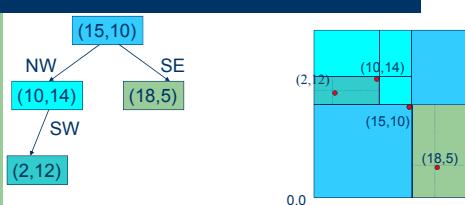
Quindi la nuova radice del sottoalbero è una delle sue foglie.

La foglia da selezionare è quella che permette di cambiare il minimo possibile la struttura dell'albero e che lo mantiene il più bilanciato possibile (esistono euristiche che lo fanno).

Esempio di euristica: per ogni foglia candidata ad essere la nuova radice si identificano due regioni, una delimitata dalle rette verticali passanti per la foglia e per il nodo da rimuovere ed una delimitata dalle rette orizzontali passanti per la foglia e il nodo da rimuovere.

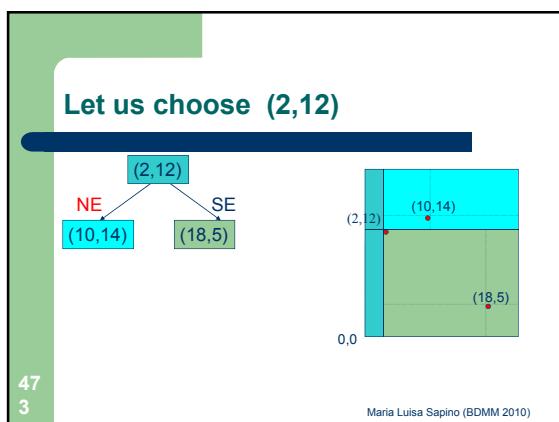
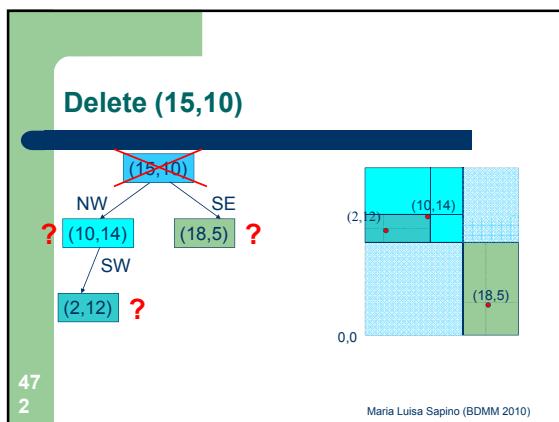
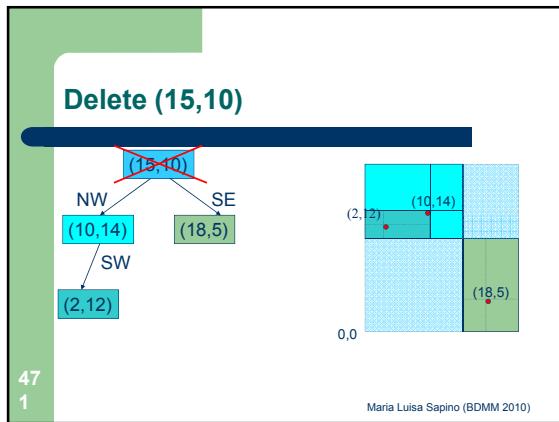
In queste fasce si trovano i punti la cui collocazione potrebbe variare nella foglia rispetto alla radice (ad esempio un punto prima era a NW della radice e ora sarebbe a NE della possibile nuova radice, cioè della foglia candidata). Quindi più punti vi sono in quell'area, più è possibile che si debbano fare delle modifiche all'albero una volta cambiata la radice e quindi che l'operazione sia più costosa. Assumendo una distribuzione uniforme del database, più piccole sono queste due fasce, minore è la probabilità che vi siano punti in esse, per cui minore è la probabilità di effettuare modifiche. Si sceglie dunque la foglia che ha le due fasce più piccole come nuova radice

Delete (15,10)



47
0

Maria Luisa Sapino (BDMM 2010)



cancellazione è caso complicato: esistono algoritmi per trovare il candidato sostituto per il nodo che si va a cancellare!

in alcuni casi può richiedere di modificare le posizioni reciproche dei punti nell'albero

in realtà non si cancellano davvero ogni volta, ma si cancellano solo "virtualmente": dopo un po' si ricostruisce direttamente l'indice senza i nodi cancellati (magari offline)

K-Nearest Neighbor Search

- This is the most important operation.
- Given a query point Q, find the K closest (to Q) points in the point quadtree.
- For simplicity, we will focus on K=1. Easy to generalize to K > 1.

Maria Luisa Sapino -
Basi di dati Multimediali

primo caso di QUERY: query top-k con k=1
ovvero

---> trovare, dato un punto nello spazio indicizzato bidimensionale, l'elemento più vicino tra quelli che si trovano nell'indice

K-NN Search

- S – a set
- A metric d: $S \times S \rightarrow \mathbf{N}$ is a mapping s.t.
 - $d(x,x) = 0$
 - $d(x,y) = d(y,x)$
 - $d(x,y) + d(y,z) \geq d(x,z)$
- We extend d to a function from $S \times 2^S \rightarrow \mathbf{N}$ as follows:
 - $d(x,R) = \min\{d(x,y) \mid y \in R\}$

Maria Luisa Sapino -
Basi di dati Multimediali

strutture indice sono significative se si lavora in uno spazio metrico, ossia la misura deve soddisfare i requisiti della metrica

- * riflessività
- * simmetricità
- * disuguaglianza triangolare

estendiamo la misura da un punto a una REGIONE:
distanza di x da R = distanza di x dal punto che in R gli è più vicino -> importante perché il pruning (potatura) viene fatto in base a questa distanza

N.B. ciascun nodo implicitamente rappresenta una regione

L'algoritmo procede per approssimazioni successive fino a trovare la soluzione.

bestSOL = soluzione migliore trovata in un determinato momento
(coincide con il risultato alla fine dell'algoritmo)

bestDIST = la distanza del punto più vicino (che è bestSOL dalla query).

Ad ogni nodo visitato: si calcola la distanza tra tale nodo e la query -> se è minore di bestDIST quel nodo diventa la soluzione migliore (aggiorna bestSOL e bestDIST).

Si esplorano i figli del nodo e si potano quelli la cui distanza tra query e il punto più vicino della regione che denotano è maggiore di bestDIST (quasi punto all'interno è maggiore di un punto già trovato per cui non è sicuramente il più vicino). Si procede esplorando uno dei figli rimanenti.

Quando non ci sono più nodi da visitare l'algoritmo termina.

bestSOL = (init) null
bestDIST(init) inf

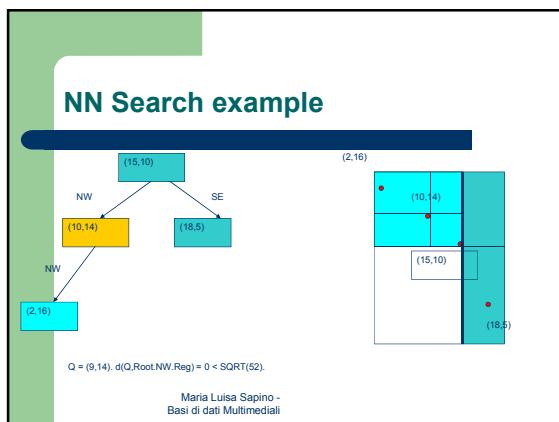
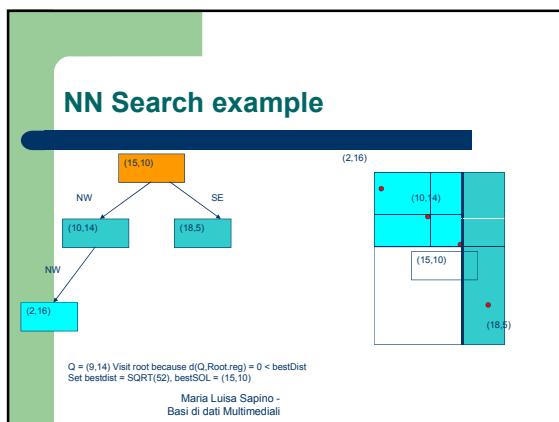
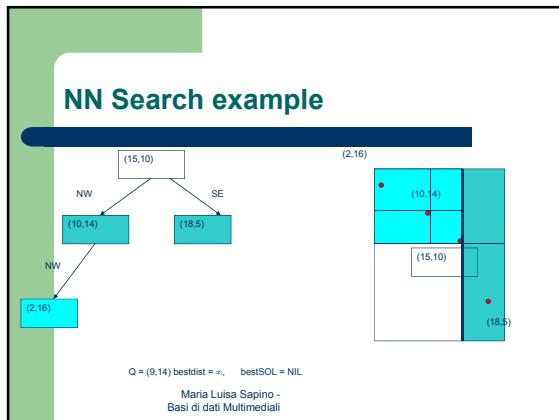
La distanza tra la query e una regione in cui si trova la query è zero.

K-NN Search

- Each node N *implicitly* represents a region N.reg.
- Algorithm for NN search works as follows.
 - Maintain variable bestdist (initialized to ∞)
 - Maintain variable bestSOL (initialized to NIL)
 - Algorithm visits nodes starting from root.
 - Everytime it visits a node N, it examines the point labeling that node. If $d(Q,N.point) < bestdist$, it updates bestdist and bestSol. Otherwise it continues.
 - Only nodes in such that $d(Q,N.reg) < bestdist$ are visited. WHY?

Maria Luisa Sapino -
Basi di dati Multimediali

QUERY: (9, 14)

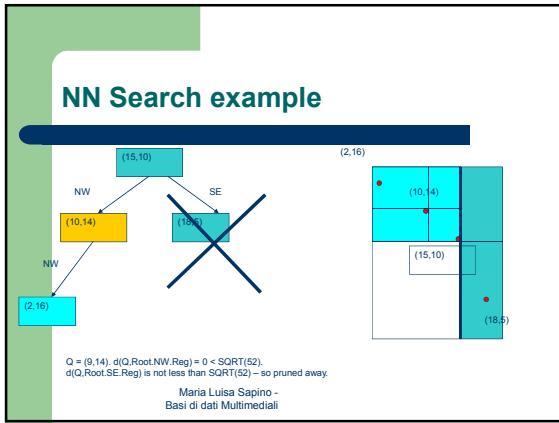


parto dalla radice:

- 1) la regione che la radice rappresenta dista dalla query meno della BESTDIST attuale? inizialmente sì perché $0 < \text{bestdist}$ -> calcolo distanza effettiva rispetto alla query (distanza euclidea tra $(9, 14)$ e $(15, 10)$) e inizializzo
- * BESTDIST = $\sqrt{52}$
 - * BESTSOL = $(15, 10)$

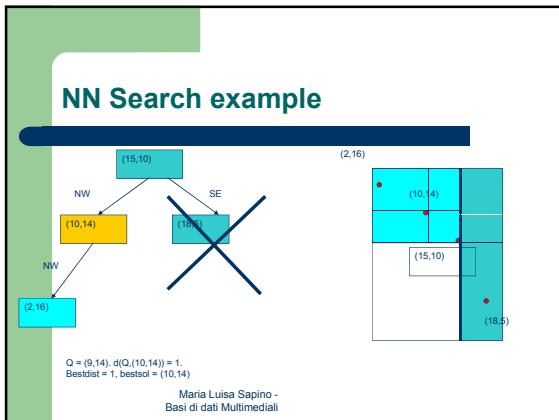
- 2) procedo andando a vedere se è il caso di continuare su figlio di dx o di sx: come? a caso oppure con euristiche, es: posso scegliere di considerare quella più vicina alla query

nell'esempio considero figlio di sx:
il nodo rappresenta la regione NW, quindi devo capire se quella regione dista dalla query più o meno rispetto alla soluzione parziale attuale: la distanza è 0 perché la query è lì dentro, quindi esploro la regione NW



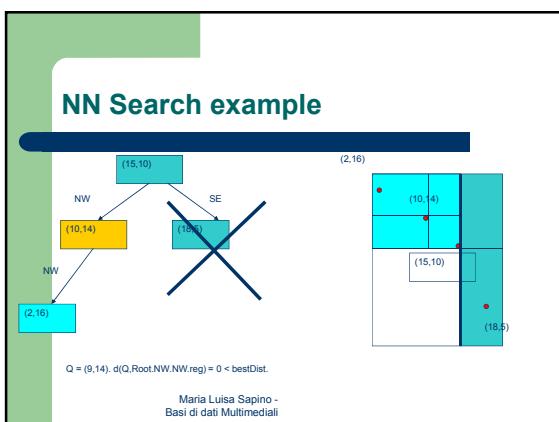
scarto il figlio di dx perché, calcolando la distanza euclidea, ottengo un valore maggiore di 0: infatti la distanza tra la query e la regione rappresentata da (18, 5) è data dalla distanza minima tra la query e i punti nella regione, ovvero la distanza tra (9, 14) e (15, 10) = $\text{sqrt}(52) > 0$

posso potare il ramo di destra perché di sicuro non trovo risultati migliori

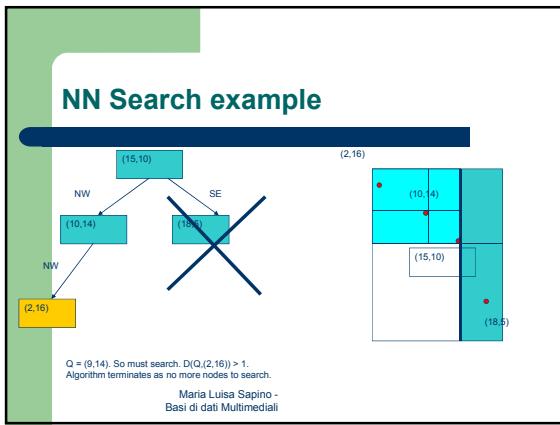


aggiorno la soluzione parziale:

- * BESTDIST = 1, cioè la distanza tra (9, 14) e (10, 14)
- * BESTSOL = (10, 14)

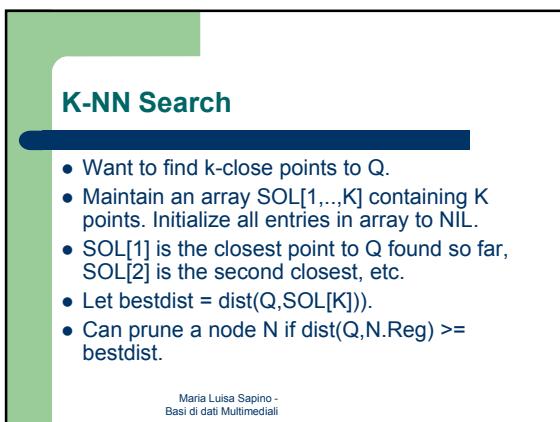


devo ancora controllare se esplorare la regione (2, 16):
la query dista 0 da questa regione, quindi la esploro e controllo la distanza, che però è > BESTDIST



non ci sono più nodi da controllare,
quindi il risultato finale è (10, 14)

==== FINE ALGORITMO DI RICERCA ===

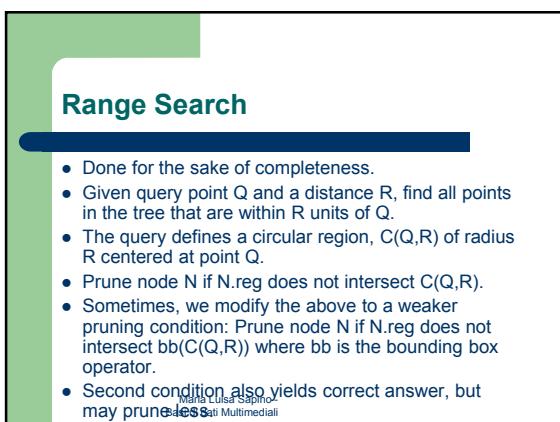


Per estendere la ricerca dell'oggetto più vicino alla ricerca dei k oggetti più vicini è sufficiente avere due array

bestSOL[1..k]: migliori k soluzioni (eventualmente array ordinato)
bestDIST[1..k]: migliori k distanze associate alle soluzioni

Per scegliere un sottoalbero è necessario che la distanza euristica tra la query e la regione delimitata dalla radice del sottoalbero sia minore del massimo valore in bestDIST, ossia che in tale regione occorra una soluzione almeno migliore della peggiore fin'ora trovata (se non è neanche migliore della peggiore allora si può potare l'albero).

Quando la distanza tra la query ed un nodo visitato è minore di un qualche valore in bestDIST, si aggiunge tale nodo e tale distanza (in bestSOL e bestDIST) eventualmente mediante un inserimento ordinato per tenere l'ordinamento dei risultati.



query di range vengono gestite in modo simile.

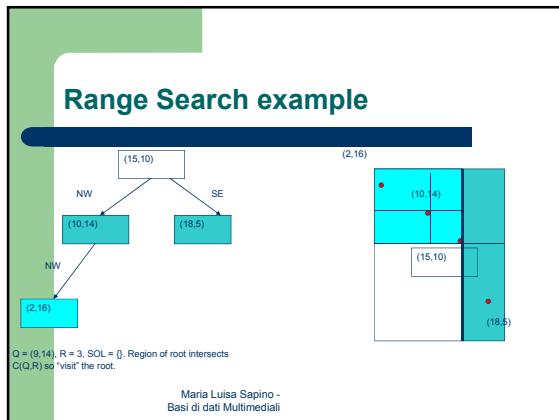
query di RANGE: dato un punto e una distanza, restituire tutti i punti nello spazio che stanno dentro a quella distanza.

* esploro tutti i nodi a partire dalla radice

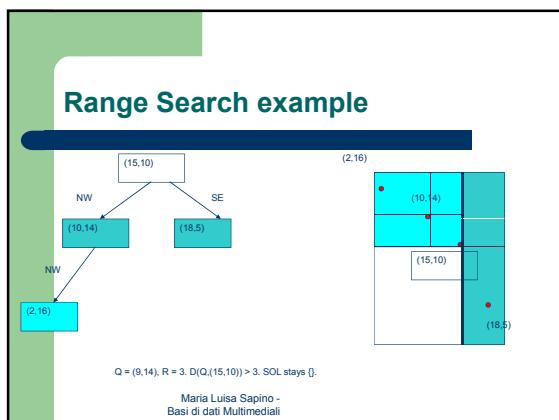
* poto se il cerchio costruito attorno alla query non interseca la regione del sottoalbero (è inutile che vada a cercare lì dentro: non posso trovare nulla che cada nel cerchio)

N.B. nella realtà si utilizza la bounding box attorno al cerchio, col rischio però di restituire di più (ma non di perdere risultati), perché è più facile calcolare le distanze

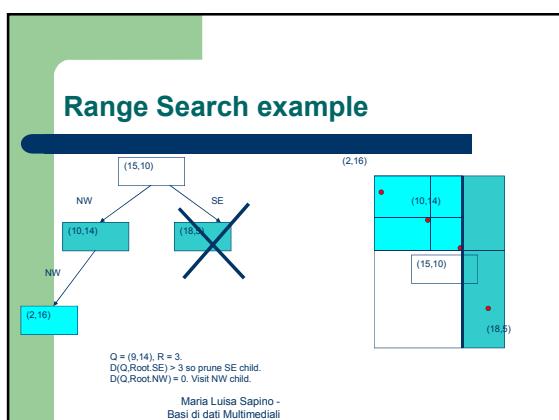
query (9, 14): trovare punti a distanza < 3



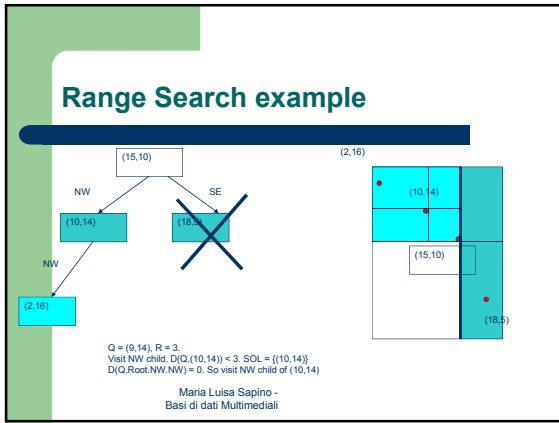
inizialmente visito la radice perché il cerchio interseca la regione associata (cioè tutto lo spazio)



visitò la radice per verificare se la sua distanza dalla query è < 3: siccome è $\sqrt{52} > 3$, non aggiorno SOL[]

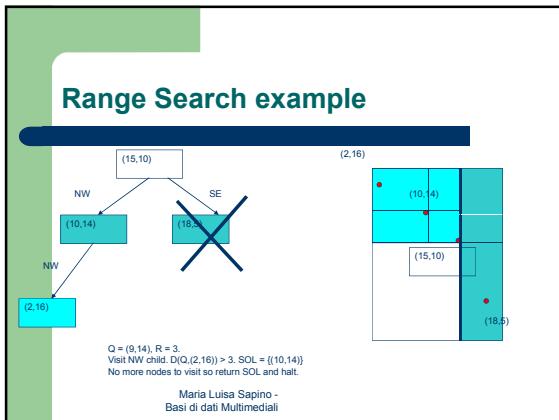


considero le due regioni associate ai figli della radice: una contiene la query e quindi ha distanza 0, l'altra ha distanza superiore a 3 quindi non viene esplorata



visit il nodo (10, 14):
poiché la distanza tra la query e il punto è compresa nel range, aggiorno SOL = {(10, 14)}

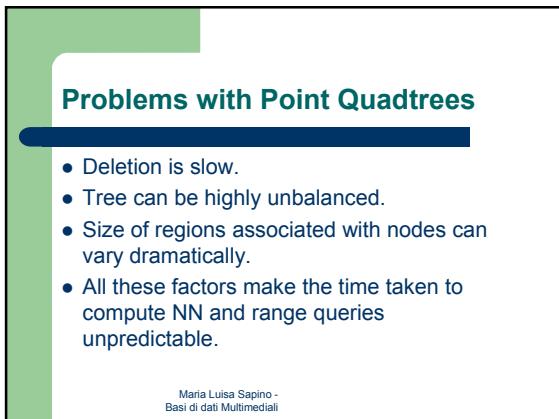
poiché la regione di (2, 16) interseca la BB della query, visito anche (2, 16)



(2, 16) ha distanza dalla query > 3, quindi non aggiungo alla soluzione.

la soluzione finale è {(10, 14)}

==== FINE ALGORITMO DI RICERCA SU RANGE ===



+) se l'albero è costruito bene (bilanciato) è efficace, consente buone potature

-) cancellazione costosa

-) albero può essere sbilanciato e se è molto sbilanciato costa più della ricerca sequenziale (ho in più il costo di gestione dell'albero!)

-) dimensioni associate alle regioni possono essere molto diverse: le regioni molto grandi aumentano il rischio di avere potature mancate

Tutto questo rende imprevedibile il tempo necessario per risolvere una query.

- le BDM sono abbastanza statiche come contenuto (non si inserisce o cancella molto)
- interrogazione su questi dati (tanti e pesanti) deve evitare ricerca sequenziale, altrimenti il tempo per il retrieval sarebbe troppo alto
- inoltre le query non sono di carattere esatto, quindi non mi fermo dopo aver trovato ad es. 1 oggetto ma continuo con la ricerca sequenziale

==> per questi motivi si definiscono opportune strutture dati per evitare ricerca sequenziale e potare parti del DB che non contengono risultati per la query

non vanno bene quelli delle BD tradizionali, sia per il tipo di ricerca che non è supportato (ricerca per range o k-NN) sia perché i dati sono multidimensionali

=====

differenza tra CLASSIFICAZIONE e CLUSTERING

- entrambi raggruppano oggetti, ma in base a criteri diversi:

* CLASSIFICAZIONE: si raggruppano oggetti che condividono certi valori di certi attributi; si possono fare classi annidate per raffinare la classificazione, es: prima suddivido autoveicoli in base al # di ruote, poi a loro volta suddivido le due classi che ho creato in base alla velocità dividendole in base ad una certa soglia di velocità: tutti quelli che stanno nella stessa classe hanno valori che stanno all'interno del range che caratterizza il cluster (e che ho usato per classificare)

* CLUSTER: insieme di individui simili tra loro, ma senza entrare nel merito del significato di "simili" -> usiamo misure di somiglianza che eventualmente non sappiamo a cosa siano dovute (ma ovviamente possiamo anche clusterizzare se conosciamo le features!), è misurata come blackbox, ad es. in base alla preferenza degli utenti, come somiglianza tra pagine web perché un utente le visita insieme

quindi

PER CLASSIFICARE entro nel merito del valore degli attributi, PER CLUSTERIZZARE invece faccio una valutazione globale basata su un valore di somiglianza che conosco già

=====

INDICI: sono strutture ad albero bilanciate -> vogliamo trovare un limite superiore al costo della ricerca, quindi arriviamo alle foglie in tempo minore rispetto a quello richiesto da una ricerca sequenziale -> la foglia punta alla pagina in cui sono contenuti i dati che cerchiamo

--> tipicamente gli oggetti nella stessa pagina sono vicini e quindi vengono restituiti senza ulteriori accessi in memoria

=====

prendiamo idee da:

- * alberi binari (point-quadtree, kd-tree)
- * B-tree (r-tree) --> fanout min e max, split, bilanciamento

MX quadtrees

- In point quadtrees, the region is split by drawing a vertical and a horizontal line through the point labeling node N.
 - In MX-quadtrees,
 - the entire space is a $2^n \times 2^n$ matrix.
 - region is split by drawing a vertical and a horizontal line through the center of the region.

Maria Luisa Sapino (BDMM 2010)

MX quadtree: ciascun nodo è associato ad una regione ripartita in altre quattro (da questo QUAD tree). La divisione dello spazio non è più guidata dall'ordine di inserimento dei nodi (come in point quadtree) ma è fissa:

ogni volta la regione viene suddivisa in 4 parti uguali.

L'ideale è avere dunque un database in cui il numero di dati è il più vicino possibile ad una potenza di 2. Inoltre i punti stanno solo sulle foglie e i nodi interni identificano le regioni.

MX quadtrees: example

Empty region



Maria Luisa Sapino (BDMM 2010)

abbiamo spazio $4 \times 4 = 2^2 \times 2^2$

per rappresentare il punto in figura, devo vedere dove si trova rispetto alla radice che rappresenta l'intera regione

poiché la suddivisione è fatta a priori, i dati sono tutti a livello delle foglie! -> la struttura non è definita dai dati che sono inseriti, ma è definita a priori
+) l'albero è sempre bilanciato e l'altezza è data dal $\log_2(\text{dimensione della matrice})$
es: in questo caso è 2

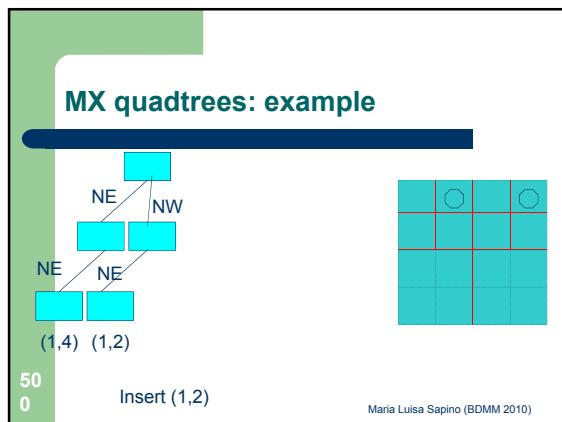
l'albero non può mai degenerare in una lista!

MX quadtrees: example



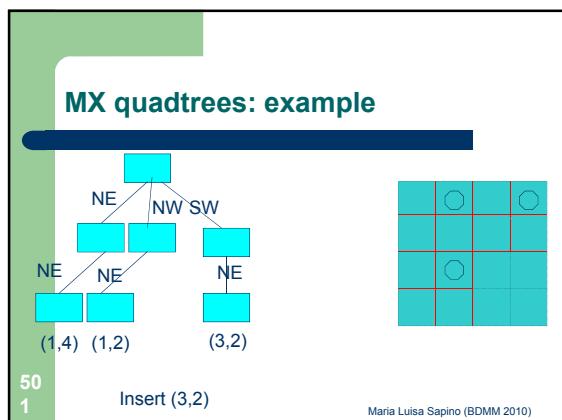
Insert (1.4)

Maria Luisa Serrano (PDMM 2010)



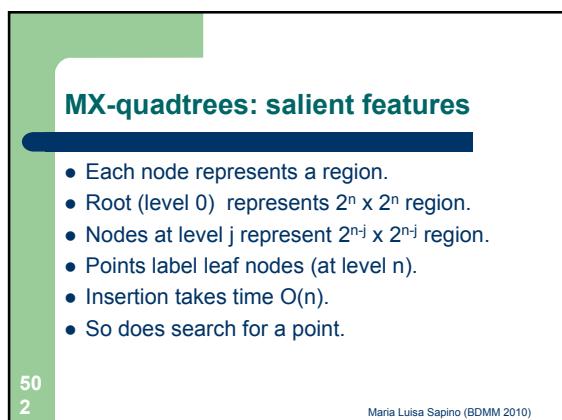
N.B. nei point-quadtrees ho punti di sistema cartesiano, qui invece ho celle di una matrice -> quindi devo fare attenzione a quali sono i valori della coppia (x, y):

* nei point-quadtrees sono valori di ascissa e ordinata
* negli MX-quadtree sono indici di una cella



Ogni nodo rappresenta una regione. I suoi 4 figli denotano le sottoregioni NW, NE, SW, SE del nodo che sono suddivise in maniera uguale. Quindi la radice identifica l'intera area, i 4 figli identificano le aree NW, NE, ecc. del database, ciascuna delle quali è 1/4 della dimensione totale del database. Lo stesso vale ricorsivamente per i figli (ogni area viene suddivisa in 4 ecc.). Le regioni rappresentate dalle foglie sono i punti, quindi gli MX-quadtree assumono uno spazio del database discreto, mentre point quadtree uno spazio continuo.

L'albero è sempre bilanciato. Per costruire l'albero si sceglie il livello n di profondità dell'albero, così da



MX-Quadtrees: deletion

- Very easy to delete a point.
- First search for the point (which must be a leaf) and delete the leaf.
- If the parent now has 4 empty child fields, then delete the parent. And repeat as long as possible. This process is termed “collapsing”.

50
3

Maria Luisa Sapino (BDMM 2010)

per cancellare un punto basta trovarlo:

1) cancello la foglia

2) se il genitore non ha altri figli cancello anche il genitore, e così via fino a che non raggiungo un genitore che ha altri figli e quindi non può essere cancellato (operazione di COLLAPSING)

VANTAGGI

- +) regioni di dimensioni omogenee
- +) ricerca e inserimento O(n)
- +) struttura semplice

SVANTAGGI

-) necessità di definire a priori la dimensione del dominio -> discretizzazione dei dati!
es: se nell'albero di prima devo inserire (2.5, 3) non posso, devo discretizzare tutto, quindi non è semplicemente applicabile

struttura che combina i vantaggi dei due approcci visti (point e MX quadtree) -> PR-quadtree, ovvero Point Region quadtree

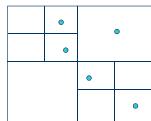
* lo spazio viene diviso in quadranti uguali (come MX-qt)

* non richiede struttura matriciale, ma consente di avere all'interno delle regioni punti distribuiti (come point-qt): quando la distribuzione dei punti in una regione supera una certa soglia, suddivido la regione

Cogniuga i vantaggi di point-quadtree e MX-quadtree.

PR-quadtrees

- MX-quadtree works well if the data is discrete
 - otherwise, it may need to use buckets, which may increase search time
- PR-quadtree (point region quadtree) assumes a continuous space.



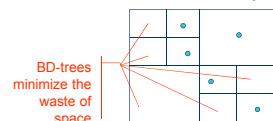
Structure is independent of insertion order
Deletion is easy

Maria Luisa Sapino (BDMM 2010)

50
6

PR-quadtrees

- MX-quadtree works well if the data is discrete
 - otherwise, it may need to use buckets, which may increase search time
- PR-quadtree (point region quadtree) assumes a continuous space.



Structure is independent of insertion order
Deletion is easy

Maria Luisa Sapino (BDMM 2010)

50
7

la cancellazione è facile, perché le direzioni non sono definite in base al nodo che inserisco, quindi non devo ristrutturare la griglia quando cancello un nodo

KD-trees

- Deficiencies of quadtree: 2^k
 - each node requires k comparisons
 - each leaf contains k null pointers
 - node size gets larger as k increases

51
0

Maria Luisa Sapino (BDMM 2010)

Nel Pont e MX quadtree, all'aumentare delle dimensioni dello spazio aumenta il numero di confronti ad ogni livello dell'albero per localizzare le varie regioni. A k dimensioni l'albero ha ampiezza 2^k , quindi ad esempio nelle query occorrono 2^k confronti e in più si spreca molto spazio in MX-quadtree (ci sono 2^k nodi interni solo al primo livello che non rappresentano foglie).

KD-trees

- Deficiencies of quadtree:
 - each node requires $\log k$ comparisons
 - each leaf contains k null pointers
 - node size gets larger as k increases
- Solution: KD-tree
 - the tree is binary whatever k is!!!
 - each node has two pointers only

51
1

Maria Luisa Sapino (BDMM 2010)

KD-tree: i punti suddividono le regioni. A livello i -esimo si suddivide secondo la dimensione i -esima. Ad ogni livello ci sono due nodi che suddividono lo spazio in due rispetto ad una dimensione.

Radicè: tutto lo spazio

- 2 Figli: un figlio ha una parte di uno spazio e l'altro l'altra parte (suddivisa rispetto alla prima dimensione e rispetto al punto dato - come in Point Quadtree)
- 4 Figli: due suddividono ulteriormente la prima parte dello spazio in due parti rispetto alla seconda dimensione e l'altro suddivide l'altra parte in due parti rispetto all'altra dimensione.
-

è un ALBERO BINARIO QUALUNQUE SIA IL NUMERO DI DIMENSIONI DELLO SPAZIO. Ma la profondità dell'albero è maggiore

K-d trees

- Used to store K -dimensional data, i.e. points of the form (x_0, \dots, x_{K-1})
- Assuming the root is a level 0 node, each node at level i discriminates on $x_{i \bmod K}$.
- Always split region associated with a node into two parts.
- We now focus on $K=2$.

Maria Luisa Sapino -
Basi di dati Multimediali

* ciascun nodo rappresenta una regione
* il punto associato ad un nodo divide la regione in due parti, ma solo rispetto ad una delle K dimensioni, che vengono considerate in modo round-robin

- 1) primo inserimento -> divido rispetto a x_0
 - 2) secondo inserimento -> divido rispetto a x_1
ecc ecc
fino a x_n , e poi riparto da x_0
-

VANTAGGI: l'albero è sempre binario per cui si fanno meno confronti per ogni livello. Si spreca meno spazio.

2-d-trees

- Structure of a node in a 2-d-tree

- nodetype =record
 - INFO: infotype; (information content. It depends on the application domain)
 - xcoord:real; ycoord:real; (coordinates of the point associated to the node)
 - Llink: @nodetype;
 - Rlink: @nodetype;
- end

Maria Luisa Sapino -
Basi di dati Multimediali

2-d-tree

- 2-d tree is a **binary tree** such that:

- If N is a node such that level(N) is even , then for every node M in the subtree rooted at N.Llink, and for every node P in the subtree rooted in N.Rlink,
 - $M.xcoord < N.xcoord \quad P.xcoord \geq N.xcoord$
- If N is a node such that level(N) is odd , then for every node M in the subtree rooted at N.Llink, and for every node P in the subtree rooted in N.Rlink,
 - $M.ycoord < N.ycoord \quad P.ycoord \geq N.ycoord$

Maria Luisa Sapino -
Basi di dati Multimediali

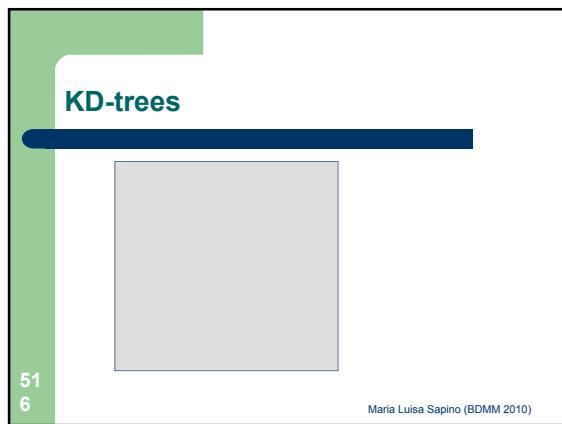
ciascun nodo è associato a una regione:
se il livello è pari discriminò rispetto a x, se è dispari
discriminò rispetto a y -> per capire se un nodo è a sx
o a dx discriminò solo rispetto alla dimensione che è
data dal livello a cui mi trovo

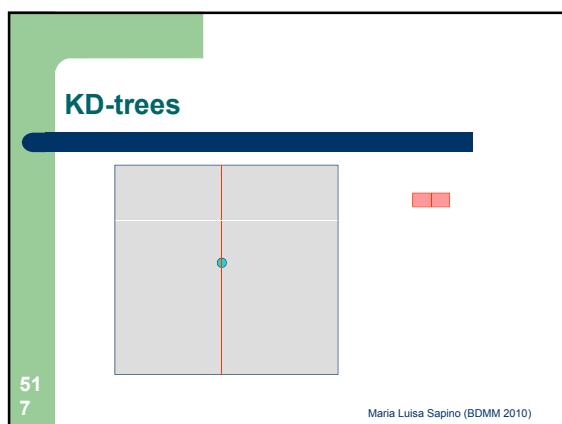
2-d-tree (notes)

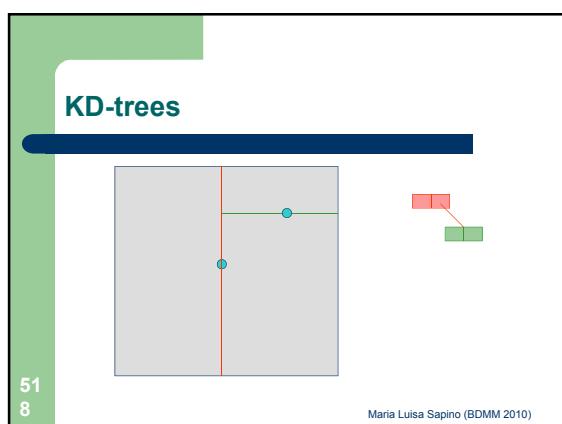
- Every node partitions the space in 2 parts:

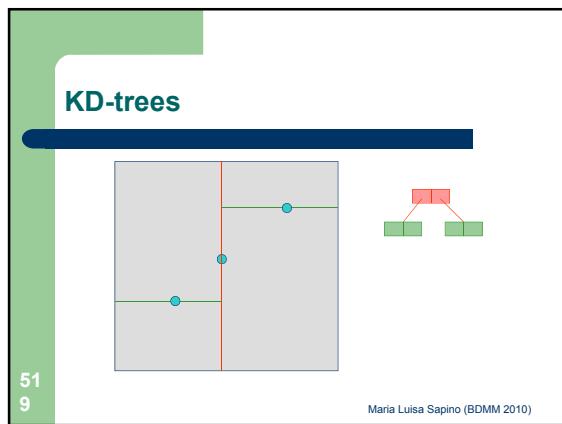
- Nodes whose level is even, implicitly draw a vertical line, $x=xcoord$,
- Nodes whose level is odd, implicitly draw a horizontal line, $y=ycoord$.

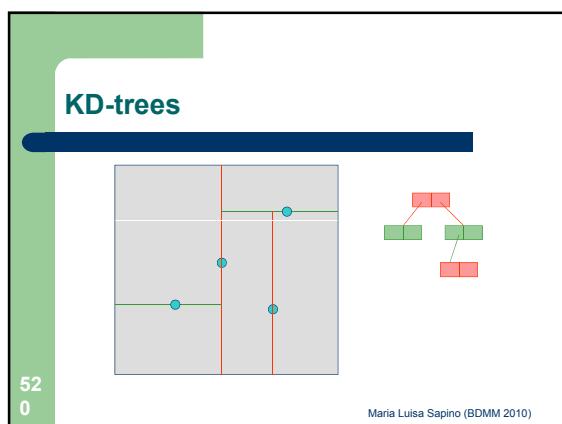
Maria Luisa Sapino -
Basi di dati Multimediali

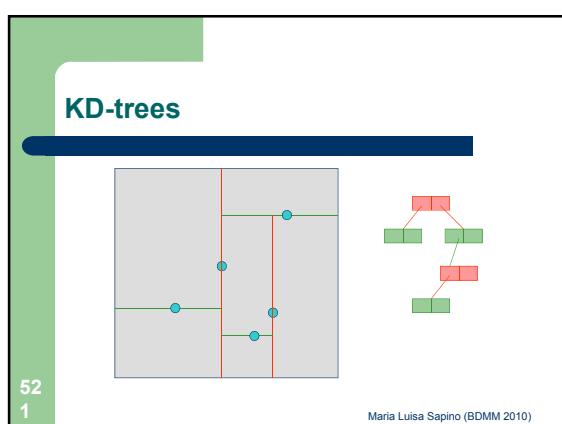












KD-trees

52
2

Maria Luisa Sapino (BDMM 2010)

KD-trees

52
3

Maria Luisa Sapino (BDMM 2010)

2-d-tree (example)

Maria Luisa Sapino -
Basi di dati Multimediali

1) prima inserisco la radice (19, 45)

- 2) a livello 0 considero solo la x:
 * $4 < 19$, quindi inserisco (4, 4) a sx
 * $40 > 19$, quindi inserisco (40, 50) a dx

3) a livello 0 considero solo la x:
 * $38 > 45$, quindi inserisco (38, 38) a dx
 siccome quel posto è già occupato, a livello 1
 considero la y:

- * $38 < 50$, quindi inserisco (38, 38) a dx

4) stessa cosa per (54, 40)

range-queries

- Given a point (x_c, y_c) and a distance d , find the set of all points (x, y) such that (x, y) lies within distance d of (x_c, y_c) .
 - Each node N implicitly represents a region R_N , constrained by N 's coordinates and its parent's coordinates.
 - If the circle specified in the query has no intersection with R_N , then there is no point searching the subtree rooted at N .
 - example: search for the circle with center (35,46) and radius 9.5 (returned: M(38, 38))

Maria Luisa Sapino -
Basi di dati Multimediali

parto dalla radice e controllo se valga la pena esplorare un sottoalbero in base al fatto che il range intersechi la regione rappresentata dal nodo radice del sottalbero

K-d-tree, k>2

- Extensions of 2-d trees, in which
 - Fields xcoord and ycoord in the definition of the node are replaced by a single field COORD, a vector of k elements.
 - For every node N , let $i = \text{level}(N) \bmod k$.
 - For every node M in N 's left subtree:
 - $M.\text{VAL}[i] < N.\text{val}[i]$
 - For every node M in N 's right subtree:
 - $M.\text{VAL}[i] \geq N.\text{val}[i]$

Maria Luisa Sapino -
Basi di dati Multimediali

...notes

- k-d trees are easy to implement
- A tree with k nodes can have height k
 - insertion and deletion can be expensive
- range searching costs, in the worst case, $O(k n^{1-1/k})$

Maria Luisa Sapino -
Basi di dati Multimediali

Points as a way of storing image data: discussion

- Each region is a rectangle with various numerical properties.
- These properties include the following.
 - Color
 - Texture
 - Location
 - And many more.....
- When using points to store image data, we reduce each rectangle to a vector. The elements of the vector describe numbers reflecting properties.

Maria Luisa Sapino
Basi di dati
Multimediali

I punti identificati nello spazio sono in realtà vettori che contengono features come istogramma di colori, di tessiture, vocabolario di un testo, ecc. e sono questi vettori che si utilizzano per indicizzare, oppure vettori per codificare coordinate spaziali per es. cartine geografiche.

Le regioni (suddivise dagli indici) possono quindi rappresentare regioni geografiche, oppure intervalli in cui ci sono certe proprietà e quindi non aree spaziali vere e proprie (ad esempio possiamo associare una dimensione ad una feature ed un'altra ad un'altra feature e le regioni sono aree in cui una feature assume certi valori e l'altra ne assume altri).

Examples: Color Histograms

- Suppose we are given an $(M \times N)$ image and suppose we are given a set $\mathbf{C} = \{C_1,..,C_k\}$ of colors.
- A *color histogram* is a mapping $ch:\mathbf{C} \rightarrow [0,1]$ such that $ch(C_i)$ is the probability that an arbitrary pixel in the image has color C_i .
- In this case, we can represent the image by a vector $\langle v_1,..,v_k \rangle$ where $v_i = ch(C_i)$.
- This vector is a *point* in a k -dimensional space. Can then use point based retrieval methods to retrieve it.

Similarity with color Histograms

- Similarity between two images is given by the distance between the color histograms of the two images.
- Very fast to compute.

Maria Luisa Sapino -
Basi di dati
Multimediali

Generalized rectangles

- Suppose we have a k -dimensional space \mathbf{R}_k .
- A generalized rectangle over \mathbf{R}_k is defined by a set of inequalities
 - $L_1 \leq x_1 \leq U_1$
 - ...
 - $L_k \leq x_k \leq U_k$
- Thus, an image can be thought of as a region in a k -dimensional space rather than a point leading to (potentially) higher fidelity.

Maria Luisa Sapino -
Basi di dati
Multimediali

R-trees

- R-trees are used to store *two* dimensional rectangle data.
- They can be easily generalized to higher dimensions.
- R-trees themselves generalize the well known B-tree.

Maria Luisa Sapino -
Basi di dati
Multimediali

R-Tree: generalizzazione del B-Tree, ma indicizza dati a due dimensioni e sono facilmente generalizzabili a più dimensioni.

Servono per indicizzare regioni piuttosto che punti (utile quando i dati sono di tipo fuzzy, es. vettori bidimensionale che identificano un colore ed una forma - un oggetto può essere di un rosso compreso tra 100 e 110 e una forma in cui non si sa esattamente se è un cerchio o tende ad essere un'ellisse). In generale le regioni identificano oggetti in maniera imprecisa.

- Ogni nodo è composto da al più N elementi (ma deve avere almeno $N/2$ elementi, come per i B-Tree).

Ogni elemento identifica una regione. L'unione di tutti gli elementi di un nodo è una regione contenuta in quella identificata dal nodo.

La radice identifica l'intero spazio e l'unione delle regioni identificate dai suoi elementi è inclusa nello spazio del DB (ma non è necessariamente l'intero spazio).

Ad ogni elemento della radice si associa un figlio che rappresenta la regione identificata da quell'elemento. Ogni figlio della radice a sua volta ha al più N elementi inclusi nella regione identificata dal figlio e così via.

Node capacity

- Each node in a R-tree can contain upto N rectangles.
- But in addition, each node must contain *at least* $N/2$ rectangles.
- We will assume henceforth that $N \geq 4$.

Maria Luisa Sapino -
Basi di dati
Multimediali

Node structure

- Each node has between $N/2$ and N rectangles.
- Like a B-tree:
 - All leaves are at the same level
 - Root has at least two children unless it's a leaf

Maria Luisa Sapino -
Basi di dati
Multimediali

- * ogni nodo ha un # di nodi compreso tra $N/2$ e N
 - * tutte le foglie sono allo stesso livello e la radice ha almeno 2 figli a meno che non sia una foglia
-
-
-
-
-

Node properties

- Each node *implicitly* represents a region.
- Root represents the whole space.
- The region of a node N , $N.\text{reg}$, is the bounding box of the rectangles stored at that node.
- Unlike quadtrees, it is possible for regions of siblings to intersect.

Maria Luisa Sapino -
Basi di dati
Multimediali

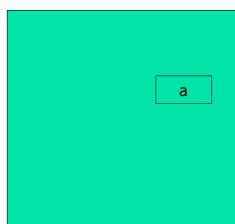
- come prima
- * ogni nodo rappresenta implicitamente una regione
 - * la radice rappresenta tutta la regione
-

cosa cambia?

- * regioni associate ai diversi nodi si adattano in base al riempimento delle regioni stesse: non sono definite a priori a partire solo dagli antenati o dal livello, ma sono definite a partire dagli elementi che stanno nella regione: sono il minimo bounding box che comprende tutte le regioni rappresentate dai figli
 - > scendendo di livello non partiziono lo spazio ma associo una MBB
 - * regioni possono intersecarsi
-

Example R-tree

a

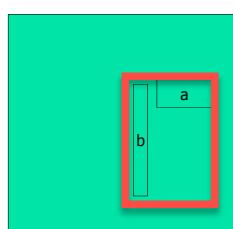


Maria Luisa Sapino -
Basi di dati
Multimediali

Radice identifica lo spazio (regione rettangolare) minimo che racchiude tutti i suoi elementi quindi in questo caso a.
Si inserisce la regione a (che nella realtà si identifica con i vertici che costituiscono la regione) come primo elemento della radice.

Example R-tree

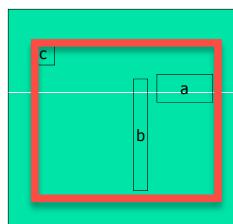
a | b | c



Maria Luisa Sapino -
Basi di dati
Multimediali

Example R-tree

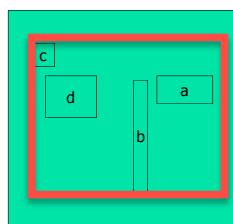
a | b | c



Maria Luisa Sapino -
Basi di dati
Multimediali

Example R-tree

a | b | c | d



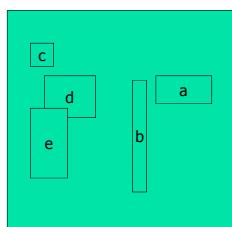
Maria Luisa Sapino -
Basi di dati
Multimediali

la regione associata alla radice è la MBB che circonda tutti gli elementi {a, b, c, d}

---> ogni nodo è associato alla MBB che racchiude tutti gli elementi che stanno nel nodo

Example R-tree

a | b | c | d



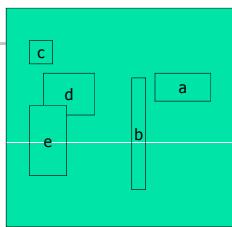
Maria Luisa Sapino -
Basi di dati
Multimediali

e non ci sta (max è N = 4):

quindi devo fare un split perché i due nodi ottenuti
contengano 3 e 2 nodi (infatti min è N/2 = 2!)

Example R-tree

a | b | c | d



Maria Luisa Sapino -
Basi di dati
Multimediali

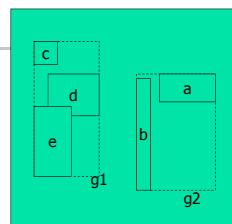
come fare lo SPLIT?

utilizzo un certo CRITERIO: devo aggregare all'interno dello stesso nodo le regioni la cui MBB ha dimensione tale che la somma delle due aree ottenute è minima

perché vogliamo MBB il più possibile compatte?
perché quando io farò una query, per decidere se
controllare un sottoalbero confronto con la MBB: se
dentro alla MBB c'è tanto spazio sprecato, è più probabile
che si controlli il sottoalbero inutilmente

Example R-tree

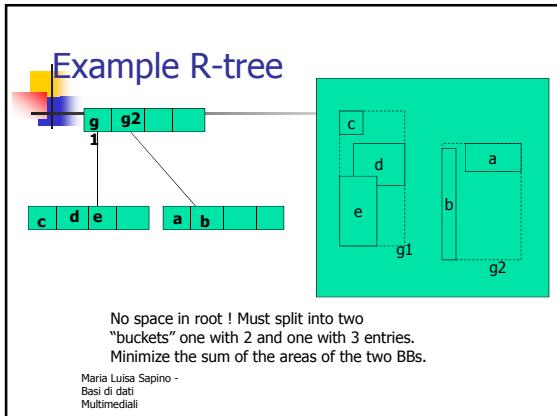
a | b | c | d



Maria Luisa Sapino -
Basi di dati
Multimediali

N.B. negli R-tree non è più vero che la regione del nodo è l'unione delle regioni dei figli!

l'unica relazione che vale è quella di contenimento: la
regione del nodo padre contiene le regioni dei nodi figli



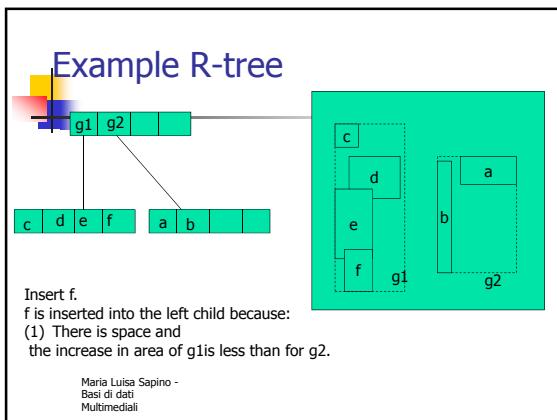
La radice contiene ora le aree g_1 e g_2 che sono le minime che ricoprono le aree identificate dai loro figli (rispettivamente c, d, e e a, b).

Splitto in base a due considerazioni

1) Tutti i nodi devono avere da $N/2$ elementi a N elementi

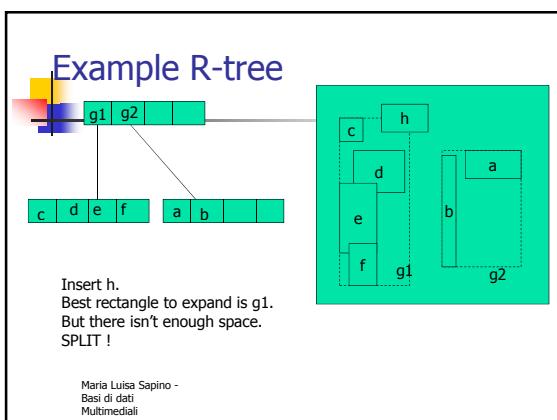
2) Le nuove regioni formate devono essere tali che

le loro aree sono più piccole possibili.



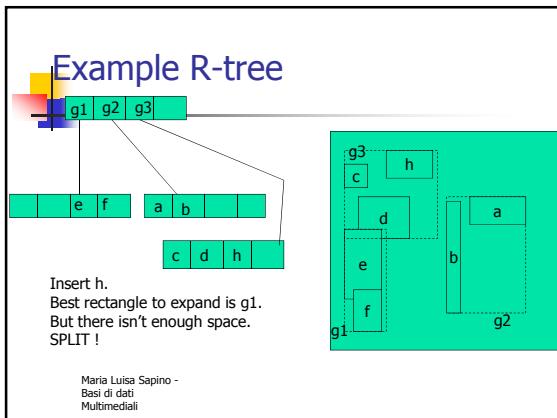
quindi metto f nella regione che fa crescere meno lo spazio complessivo.

Siccome nell'esecuzione delle query si analizza la distanza tra queste e i punti più vicini delle varie regioni da considerare, più la regione è grossa più questa euristica sarà imprecisa (posso ad es. rilevare aree molto lontane come più vicine rispetto ad altre) e quindi si pote di meno e le query vengono risolte meno efficientemente.

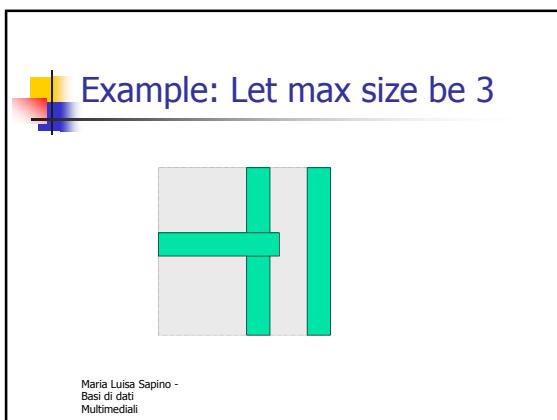


dovrei mettere h in g_1 , ma non ci sta!

soluzione: SPLITTING
introduco nuovo nodo g_3

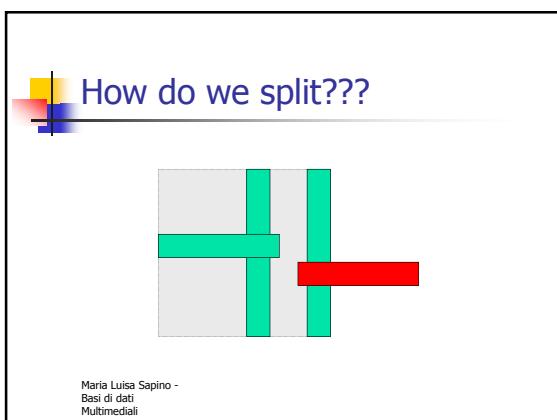


devo togliere due elementi da g1 e metterli in g3: di nuovo, scelgo quelli che minimizzano la somma delle MBB



vogliamo anche minimizzare l'OVERLAPPING, altrimenti è possibile che si debba cercare in più cammini: questo rende la ricerca meno efficiente!

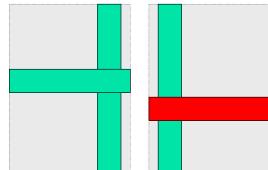
abbiamo 3 elementi all'interno di una regione



qui seguendo il criterio dato prima, dovrei unire le due regioni orizzontali e le due verticali:
in certi casi però è meglio minimizzare l'overlapping

OVERLAPPING infatti dà luogo a spazio sprecato, ovvero spazio su cui non è possibile fare pruning:
voglio regioni che siano rappresentative per il loro contenuto

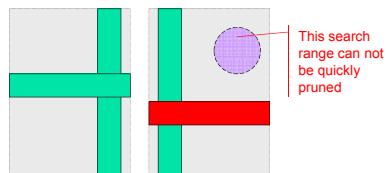
How do we split???



Maria Luisa Sapino
Basi di dati
Multimediali

Minimize overlap of the BRs

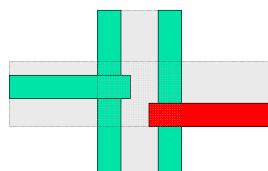
How do we split???



Maria Luisa Sapino
Basi di dati
Multimediali

Minimize overlap of the BRs

How do we split???

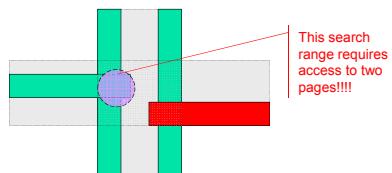


Maria Luisa Sapino -
Basi di dati
Multimediali

Minimize total area

soluzione che minimizza l'area totale non minimizza
l'overlapping

How do we split???

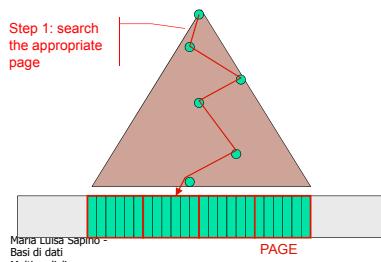


Maria Luisa Sapino -
Basi di dati
Multimediali

Minimize total area

con questa query di range, per arrivare a questo punto
dovrei seguire entrambe le strade -> arrivo allo stesso punto
da due strade diverse, quindi una delle due è inutile!

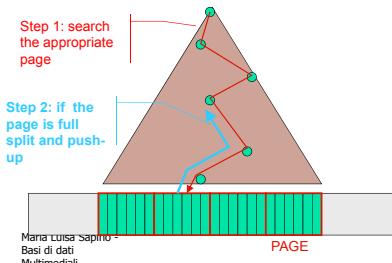
Insertion (similar to B-trees)



Maria Luisa Sapino -
Basi di dati
Multimediali

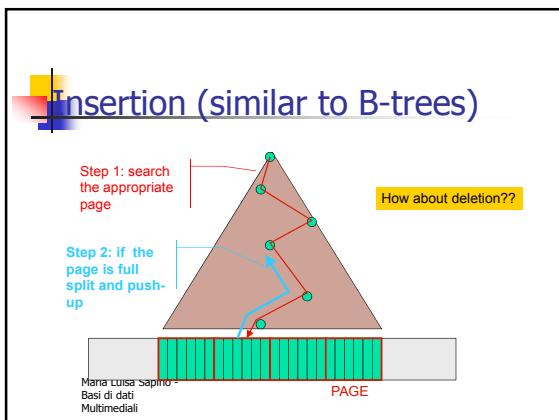
ciascun nodo ha una capienza dimensionata in base ad
una pagina di memoria (la lettura avviene leggendo
un'intera pagina ed è costosa)

Insertion (similar to B-trees)

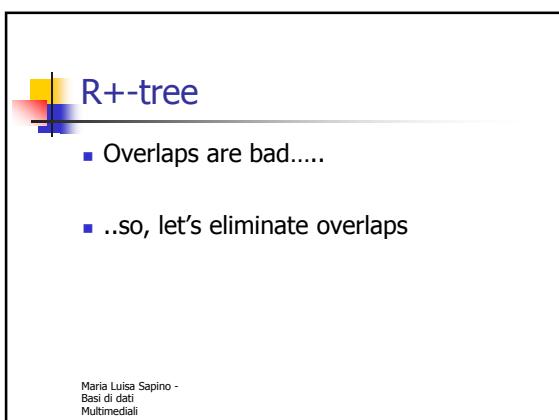


Maria Luisa Sapino -
Basi di dati
Multimediali

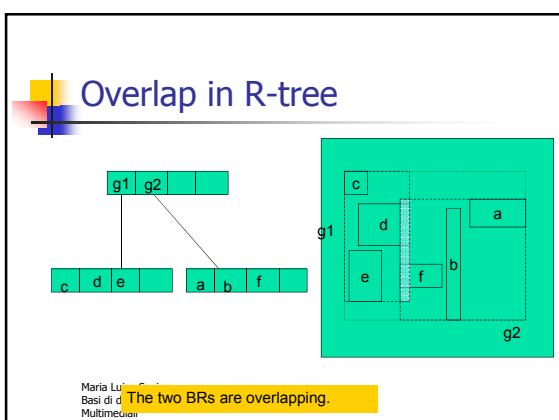
se la pagina è già piena faccio split e aggiorno il nodo
genitore



la cancellazione può comportare ristrutturazione dell'albero se un nodo diventa troppo vuoto



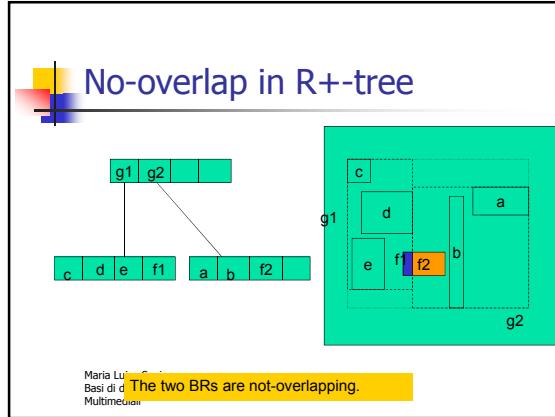
R+ tree: come R-tree ma minimizzano l'overlapping



per minimizzare l'overlapping devo gestire il caso dell'inserimento di f:
nell' R+-tree, la regione che causerebbe l'overlapping viene suddivisa in due parti, ciascuna delle quali viene associata ad una regione diversa

ovviamente bisogna poi ricordarsi che f è divisa in f1 ed f2

Si impedisce overlapping. Ad esempio f ricade sia nell'area g1 che nell'area g2. Allora si spezza f in due parti



f1: parte di 'f' che ricade nell'area 'g1'
f2: parte di 'f' che ricade nella regione g2

Poi si mantiene l'informazione che f1 non è una regione intera ma va considerata come $f = f1 \cup f2$ (e lo stesso vale per f2).

- Oppure metto f sia da una parte che da un'altra ma la bounding box di g1 rimane la stessa (cioè ricopre solo f1) e quella di g2 ricopre solo f2. In questo modo si introducono bounding box imprecise ma le query non visitano due volte una regione per trovare la stessa. Ad es. se query è più vicina a f1 che a f2 non visita g2 ma solo g1.

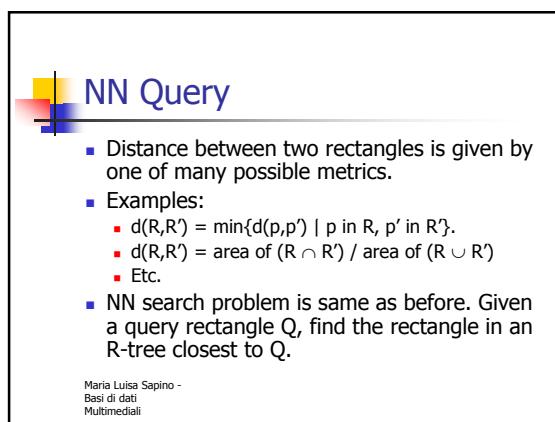
splitting basato su minimizzazione dell'area totale delle MBB è esponenziale: devo provare tutte le combinazioni e scegliere la migliore!

esiste un'alternativa più efficiente (quadratica), che è un'euristica molto meno costosa ma che dà risultati efficienti, cioè ha un rapporto costi/benefici più che accettabile

EURISTICA:

- * si trovano i due rettangoli che danno luogo alla massima bounding box -> quelli sicuramente andranno in due BB distinte
 - * distribuisco gli altri rettangoli casualmente o con un'euristica.
- NB: non necessariamente si trova la soluzione migliore ma si ha comunque un ottimo rapporto costo/benefici (si trova comunque una soluzione buona in tempi relativamente brevi, piuttosto che trovare la soluzione precisa con algoritmi intrattabili).

QUERY: dato un rettangolo, trova i rettangoli nell'indice che sono più vicini. La strategia è analoga a quella per Point tree ma si considerano rettangoli.



per fare le query bisogna prima definire una metrica:
come misurare la distanza tra due regioni?

1) distanza tra i due punti delle due regioni più vicini

2) area dell'intersezione / area dell'unione

dice quanto due oggetti sono simili
distanza = $1 - (\text{area inters.} / \text{area unione})$

quale misura usare?

dipende dal senso della regione:

es. per regione geografica è più adatta la prima metrica, mentre per rappresentazione fuzzy di un'informazione è meglio la seconda (misura quanto hanno in comune due regioni)

NN Search Algorithm

- NN search algorithm is same as before.
- Prune node N if $d(Q, N.\text{reg})$ exceeds the best distance found so far.
- Initialize bestdist and bestSol as before.

Maria Luisa Sapino -
Basi di dati
Multimediali

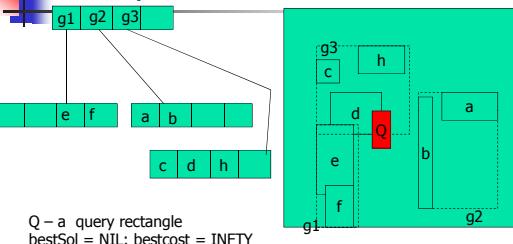
come prima:

-
- * scendo dalla radice e visito i sottoalberi
 - * poto i rami la cui regione (BB) dista dalla query più di quanto non disti il miglior risultato già ottenuto nell'esplorazione
-

inizialmente

BESTDIST = +infinito
BESTSOL = null

Example R-tree

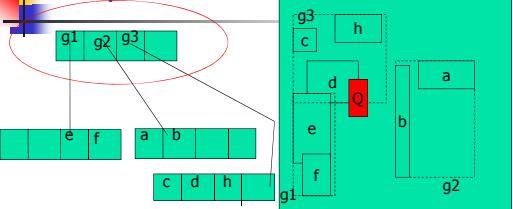


Maria Luisa Sapino -
Basi di dati
Multimediali

-
- * qui scelgo il figlio g3 perché è quello che dista meno da Q (infatti Q è parzialmente contenuta in g3)
-

* d dista 0 e quindi è soluzione

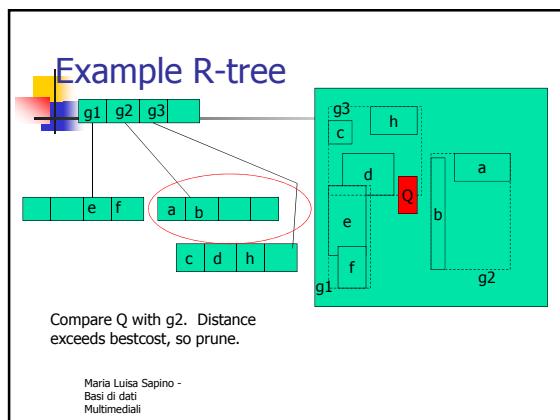
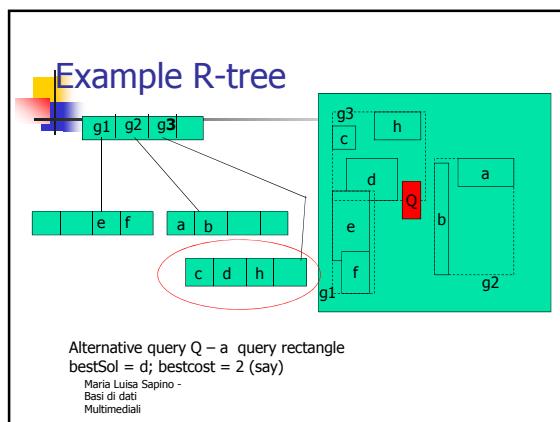
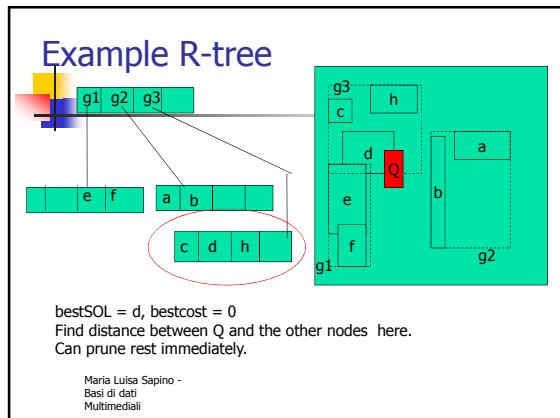
Example R-tree

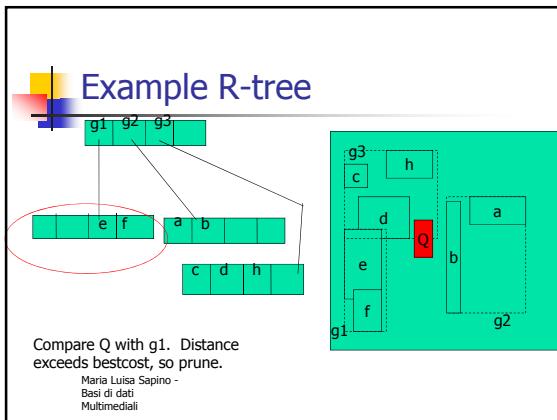


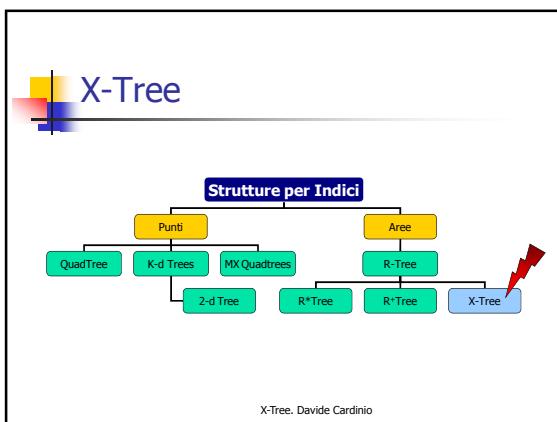
bestSOL = NIL, bestcost = INFTY
All three rectangles g1,g2,g3 to be considered. Explore G3 first as it is closest to Q.

Maria Luisa Sapino -
Basi di dati
Multimediali

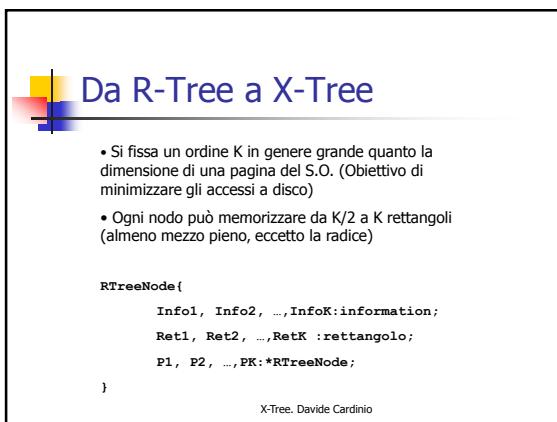
-
-
-
-
-
-
-
-
-
-



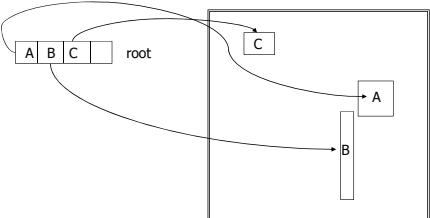




X-Tree: estensione degli R-Tree in cui i nodi non hanno capacità (numero di elementi) fissa.
 Servono perché, siccome in alcuni casi l'overlapping non è evitabile, in questi casi negli X-Tree si raddoppia la dimensione del nodo.



Da R-Tree a X-Tree



Fissato K=4

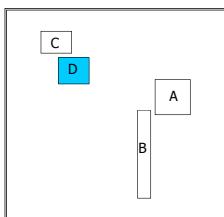
Esempio su rettangoli (2 dimensioni)

X-Tree. Davide Cardinio

Da R-Tree a X-Tree

Inserisco "D"

A B C D root



Fissato K=4

Esempio su rettangoli (2 dimensioni)

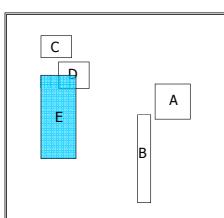
X-Tree. Davide Cardinio

Da R-Tree a X-Tree

A B C D root



Dove metto "E" ?

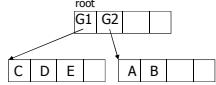


Fissato K=4

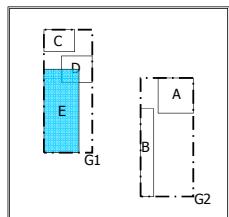
Esempio su rettangoli (2 dimensioni)

X-Tree. Davide Cardinio

Da R-Tree a X-Tree



Fissato K=4



Esempio su rettangoli (2 dimensioni)

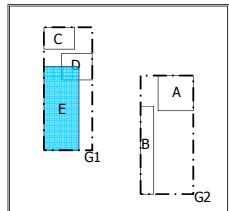
X-Tree. Davide Cardinio

1° Problema: lo split

Come faccio lo split?

Non a caso: si vogliono minimizzare gli sprechi durante la ricerca.

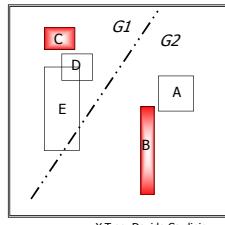
- evitare di finire dentro un rettangolo inutilmente
- problema esponenziale
- ci accontentiamo di heuristiche



X-Tree. Davide Cardinio

Split in R-Tree

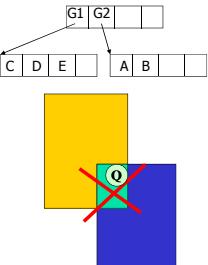
- scelgo i 2 rettangoli più lontani e li utilizzo per formare 2 gruppi di rettangoli ad essi vicini



X-Tree. Davide Cardinio

Split in R*-Tree

- Minimizza la funzione che considera
 - Area MBB (Minimum Bounding Box)
 - Aree di overlapping
 - Perimetri dei rettangoli
- Si riducono così le sovrapposizioni di aree nei rettangoli di raggruppamento dello stesso livello → ricerca nell'albero più veloce



X-Tree. Davide Cardinio

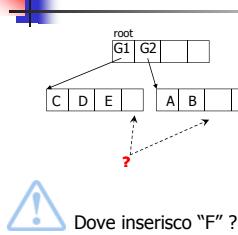
R*-tree minimizza contemporaneamente 3 grandezze:

* somma delle MBB

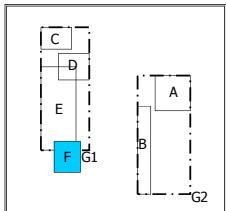
* aree di overlapping

* perimetri dei rettangoli

2º Problema: inserimento in un MBB esistente



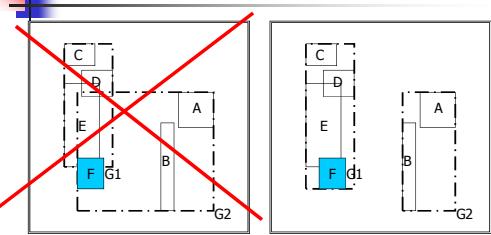
Fissato K=4



Esempio su rettangoli (2 dimensioni)

X-Tree. Davide Cardinio

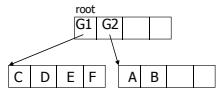
Soluzione in R-Tree



Scelgo il rettangolo che cresce di meno dopo un eventuale inserimento

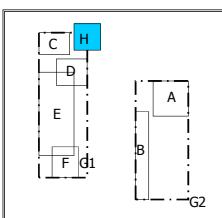
X-Tree. Davide Cardinio

3° Problema: overlapping fra MBB



 Dove inserisco "H" ?

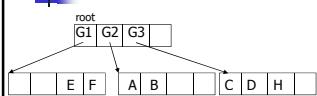
Fissato K=4



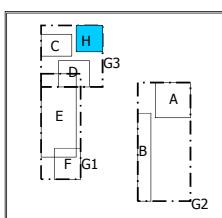
Esempio su rettangoli (2 dimensioni)

X-Tree, Davide Cardinio

3º Problema: overlapping fra MBB



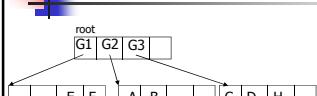
1. Split G1 in G1 e G3
 2. Inserisco "H" in G3



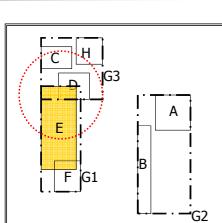
Esempio su rettangoli (2 dimensioni)

X-Tree Davide Cardinio

3º Problema: overlapping fra MBB



"E" finisce sia in G3 che in G1

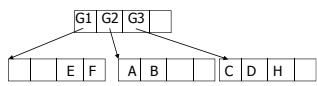


Esempio su rettangoli (2 dimensioni)

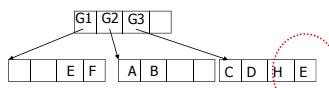
Esempio su

Overlapping fra MBB: soluzioni

- **R-Tree:** lascio le cose come stanno



- **R+Tree:** considero "E" presente in entrambi i nodi (duplicazione delle informazioni ma vantaggi per la ricerca)



X-Tree. Davide Cardinio

Da R-Tree a X-Tree

- Vantaggi degli R-Tree:
 - Spesso occorre indicizzare spazi piuttosto che punti
 - Algoritmi efficienti
- Svantaggi:
 - Deterioramento delle performance su tante dimensioni a causa degli overlaps

X-Tree. Davide Cardinio

problema: OVERLAP causa deterioramento di performance -> al crescere del # delle dimensioni, aumenta significativamente la probabilità che ci sia overlapping

All'aumentare della dimensione dello spazio aumenta la probabilità di overlapping. Già con 5 dimensioni si ha un aumento del 90 % degli overlapping.

SOLUZIONE:
introduzione degli X-tree

X-Tree

(Berchtold, Keim, Kriegel, VLDB '96)

- Struttura basata su R-tree
- Molto efficiente per tante dimensioni
- Evita l'overlap tra i nodi utilizzando
 - una strategia di split senza overlap
 - il concetto di super-nodi

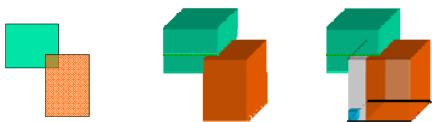
X-Tree. Davide Cardinio

poiché l'overlapping è critico, laddove lo splitting di un nodo in due non riesca ad evitare overlapping, allora si evita lo SPLIT incrementando la capacità del nodo

questo riduce il potere discriminante di un nodo nella visita, perché diventa una ricerca più sequenziale; se però io ho overlapping devo in ogni caso fare più accessi in memoria, quindi può essere più conveniente fare ricerca sequenziale in un supernodo piuttosto che seguire più rami (facendo cmq più accessi in memoria!)

Gli overlaps

- Gli overlaps si amplificano con il crescere delle dimensioni
- Incremento di oltre il 90% su 5 dimensioni

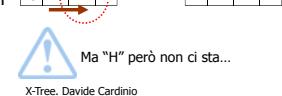
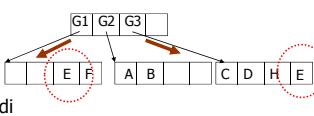


X-Tree. Davide Cardinio

X-Tree

(Berchtold, Keim, Kriegel, VLDB '96)

- Negli R-Tree un overlap implica che durante la visita, si scenda a visitare più di un figlio.
- Se debbo visitare il 90% dei figli tanto vale fare una ricerca su un albero! Uso un vettore...



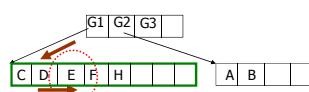
Ma "H" però non ci sta...

X-Tree. Davide Cardinio

idea del SUPERNODO: vogliamo evitare di dover scendere (nel 90% dei casi!) su più rami -> è meglio aggiungere direttamente una pagina di memoria

Il supernodo

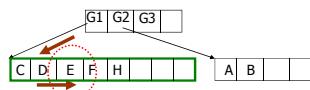
- Introduco il concetto di Supernodo, cioè un nodo di dimensioni multiple di K.
- Un X-Tree è un ibrido fra una struttura gerarchica ed una lineare



X-Tree. Davide Cardinio

Il supernodo

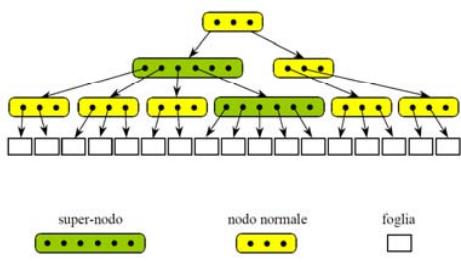
- Usato solo quando non si ha modo di evitare l'overlap



X-Tree. Davide Cardinio

se si cancella un elemento di un supernodo, se la capacità lo permette si torna alla capacità di partenza (1 pagina): in questo modo l'aggiunta di un eventuale elemento che non causerà overlapping viene fatta normalmente -> è sempre preferibile perché evita ricerca sequenziale

Esempio di X-Tree



X-Tree. Davide Cardinio

Algoritmo di inserimento

- Obiettivo: evitare gli split che producono overlap
- Nuove strategie di split
- Si determina prima il MBB dove inserire il nodo e si applica ricorsivamente l'inserimento

X-Tree. Davide Cardinio

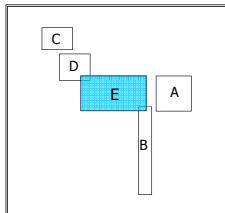
prima si tenta l'inserimento senza overlapping: se non è possibile allora si crea un supernodo

Algoritmo di inserimento

A B C D root



Dove metto "E" ?



Fissato K=4

Esempio su rettangoli (2 dimensioni)

X-Tree. Davide Cardinio

Algoritmo di inserimento

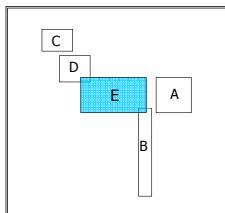
Dove metto "E" ?

A B C D

Esiste uno split
senza overlap?
No.

→ Supernodo

A B C D E | |



→ Altrimenti proseguo come per gli R*-Tree

X-Tree. Davide Cardinio

Algoritmo di inserimento

```
int X_DirectoryNode::insert (DataObject obj, X_Node **new_node)
```

```
{
```

```
SET_OF_MBR *s1, *s2;  
X_Node *follow, *new_son;  
int return_value;
```

```
follow = choose_subtree(obj);
```

Scelgo in quale
sottoalbero andare

```
return_value = follow->insert(obj, &new_son);
```

Inserisco il nodo

```
update_mbr(follow->calc_mbr());
```

Aggiusto l'area
del rettangolo del
nodo corrente

```
if (return_value == SPLIT){
```

```
...  
return SPLIT oppure SUPERNODE
```

```
} else if (return_value == SUPERNODE){
```

```
...  
}
```

```
}  
return NO_SPLIT;
```

X-Tree. Davide Cardinio

Algoritmo di inserimento

```

if (return_value == SPLIT){
    add_mbr(new_son->calc_mbr());
    if (num_of_nbrs() > CAPACITY){
        if (split(mbrs, s1, s2) == TRUE){
            set_nbrs(s1);
            *new_node = new X_DirectoryNode(s2);
        }
        return SPLIT;
    }
    else {
        *new_node = new X_SuperNode();
        (*new_node)->set_mbrs(mbrs);
        return SUPERNODE;
    }
}
...

```

X-Tree. Davide Cardinio

Algoritmo di inserimento

- I supernodi sono creati solo quando non esistono alternative
- La dimensione di un supernodo cresce di K in K
- Si cerca di tenere i supernodi in memoria
- Si può introdurre un MAX_OVERLAP calcolato in base al tempo che passa fra leggere un supernodo di lunghezza 2K e quello di leggere 2 nodi di lunghezza K

X-Tree. Davide Cardinio

N.B. qui K è la capacità di un nodo (pagina di memoria)

si può introdurre un correttivo a questa strategia "drastica" -> si tiene una soglia di overlapping consentito: MAX_OVERLAP -> se overlapping è piccolo è più improbabile che ci finisca una query

Altri algoritmi

- La ricerca "Nearest neighbor queries" simile a quella degli R-Tree con piccole modifiche per l'accesso ai supernodi
- La cancellazione deve tenere conto degli underflow dei supernodi: un supernodo di lunghezza 2K che perde il suo K+1esimo elemento deve essere convertito in un nodo normale



X-Tree. Davide Cardinio

* ricerca NN uguale a quella degli R-tree (in più bisogna solo gestire i supernodi facendo ricerca sequenziale sui supernodi)

* per la CANCELLAZIONE bisogna tenere conto degli underflow dei supernodi e compattarli quando necessario

Altri algoritmi

- L'update può essere visto come combinazione di cancellazione e inserimento

X-Tree. Davide Cardinio

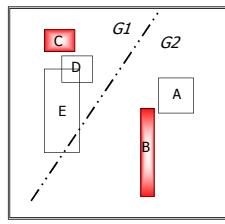
Il problema degli split

- Già risolto per 2 dimensioni
- Come faccio a renderlo efficiente per N dimensioni?
- Gli X-Tree non vogliono gli overlap

X-Tree. Davide Cardinio

Split in R-Tree

- scelgo i 2 rettangoli più lontani e li utilizzo per formare 2 gruppi di rettangoli ad essi vicini

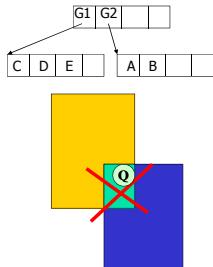


X-Tree. Davide Cardinio

Split in R*-Tree

- Minimizzo la funzione che considera

- Area MBB (Minimum Bounding Box)
- Aree di overlapping
- Perimetri dei rettangoli



- E' sempre possibile evitare gli overlap? No...

X-Tree. Davide Cardinio

Teorema

- Per punti distribuiti uniformemente, uno split senza overlap esiste se e solo se esiste una dimensione secondo la quale tutte le MBB nel nodo sono già state precedentemente splittate

X-Tree. Davide Cardinio

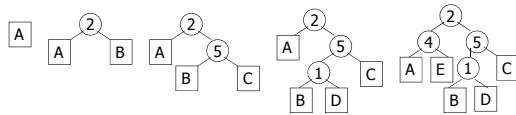
se i punti sono distribuiti in modo uniforme nello spazio, allora esiste uno split senza overlapping se e solo se la regione che si deve spartire deriva dallo spazio iniziale in seguito a molteplici split che sono stati fatti sempre rispetto alla stessa dimensione

--> in pratica si è sempre "affettato" lo spazio di partenza lungo lo stesso asse: così facendo si evitano gli incastri che sono fonte di overlapping

se invece ho affettato in più dimensioni cresce la probabilità che ci siano overlapping

Conseguenze: lo split Tree

- Per avvalerci il più possibile di questo teorema dobbiamo ricordarci la storia degli split del nodo

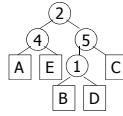


X-Tree. Davide Cardinio

definiamo struttura parallela all'indice in cui registriamo la dimensione rispetto a cui si è fatto lo splitting: la visita dei nodi interni dalla radice tiene traccia della storia delle dimensioni rispetto a cui ho fatto gli inserimenti

Utilizzo dello split Tree

- Esempio: voglio sapere come è stato ottenuto lo split di "C". Visitando l'albero dalla radice scopro che è stata prima sezionata la dimensione 2 e poi la 5



X-Tree. Davide Cardinio

In questo caso non ho garanzia di non avere overlapping (C è stato ottenuto splittando rispetto alla dimensione 2 e poi 5)

Problemi

- Non si utilizzano le dimensioni non ancora considerate nell'albero
- Sebbene una partizione esista sempre, continuando ad "affettare" sulle stesse dimensioni posso arrivare a nodi sbilanciati. Cosa che capita spesso se le dimensioni sono tante.

X-Tree. Davide Cardinio

utilizzare sempre la stessa dimensione ha alcuni svantaggi:
-> rende importante la scelta della prima dimensione su cui splittare
-> splittando sempre lungo la stessa direzione si può arrivare ad alberi sbilanciati

Se si suddivide sempre rispetto alla stessa dimensione non ho rischio di overlapping però ho un albero che è praticamente monodimensionale (nei query si individua la porzione di spazio in cui la dimensione rispetto alla quale si suddivide ha un determinato range ma a questo punto per uno spazio a k dimensioni, occorre fare una ricerca sequenziale a praticamente k-1 dimensioni).

Occorre un tradeoff tra l'utilizzo del teorema (fare splitting sulla stessa dimensione) e la costruzione di alberi le cui operazioni diventano inefficienti.

Soluzione

- Provo uno split topologico (R*-Tree)
- Se crea degli overlap provo a trovare uno split senza overlap sfruttando lo split tree
- Se il risultato è troppo sbilanciato creo un supernodo

X-Tree. Davide Cardinio

soluzione: EURISTICA

* prima provo split topologico, es. r*-tree in cui scelgo i due rettangoli la cui MBB è massima e raggruppo gli altri in base a questi due

* se anche così non riesco ad evitare l'overlapping uso lo split tree visto appena sopra, cioè divido sempre rispetto alla stessa dimensione.

* se il risultato è troppo sbilanciato -> supernodo

La procedura di split

```

bool X_DirectoryNode::split(SET_OF_MBR *in, SET_OF_MBR *out1, SET_OF_MBR *out2)
{
    SET_OF_MBR t1, t2;
    MBR r1, r2;

    topological_split(in, t1, t2);
    r1 = t1->calc_mbr(); r2 = t2->calc_mbr();
    if (overlap(r1, r2) > MAX_OVERLAP)
    {
        overlap_minimal_split(in, t1, t2);
        if (t1->num_of_mbrs() > MIN_FANOUT || t2->num_of_mbrs() < MIN_FANOUT)
            return FALSE;
    }
    *out1 = t1; *out2 = t2; return TRUE;
}

```

X-Tree. Davide Cardinio

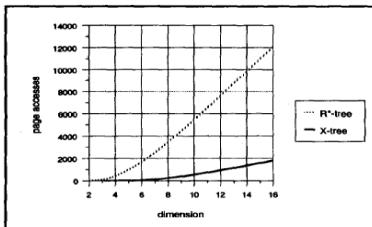
Provo lo split topologico

*Se fallisce
dando overlap
inaccettabile...*

Cerco lo split senza overlap

*Se gli split sono troppo
sbilanciati fallisco: supernodo*

Conclusioni

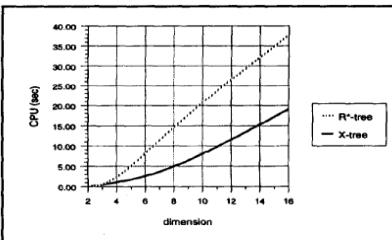


X-Tree. Davide Cardinio

grafico che correla il crescere delle dimensioni dello spazio con l'accesso alle pagine di memoria:

X-tree riduce notevolmente l'accesso alla memoria!

Conclusioni



X-Tree. Davide Cardinio

c'è anche notevole risparmio di CPU

TV-tree: prende idee da
 * indici ad albero
 * tassonomie di classificazione

TV trees (telescopic vector trees)

(Lin, Jagadish, Faloutsos, VLDB Journal, 1994)

- Based on classification idea
- Dimensionality curse: R-trees do not work for large numbers of dimensions
- Idea:
 - not all features are equally important
 - order features based on importance (discrimination power)
 - use as little features as possible
 - "contract" and "extend" feature vectors based on need

56
4

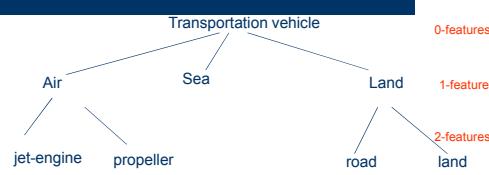
Maria Luisa Sapino (BDMM 2010)

Durante l'indicizzazione di dati, può essere ridondante rappresentare tutte le feature scelte a tutti i livelli.

L'idea dei TV-Tree è quella di non rappresentare tutte le feature in tutti i livelli (qualche livello conterrà regioni descritte su meno dimensioni di quelle totali che costituiscono lo spazio).

Il problema dell'overlapping sparisce perché il rischio diminuisce grazie all'utilizzo di meno dimensioni (anche se non a tutti i livelli)

Intuition



Classification requires less features at the higher levels than it uses at the lower levels

Maria Luisa Sapino (BDMM 2010)

posso caratterizzare i nodi con features diverse ai vari livelli, nell'es:

- al 1° livello il terreno su cui si sposta il veicolo
- al 2° il tipo di motore
- ecc...

il # di features usate cresce man mano che ci si allontana dalla radice

nei TV-tree utilizzo questa idea: costruisco struttura ad indice in cui ciascun nodo rappresenta una BB che definisco cercando di usare meno features possibile.

Scendendo di livello aggiungo sempre più feature partendo da quella più discriminante (mi permette di definire più regioni in cui ricadono molti dati) fino alla meno discriminante.

Questo tipo di alberi va bene quando si conosce già il DB per poter riconoscere che ci sono feature più importanti rispetto ad altre.

Cost of a dimension

- Every rectangle has to have values describing all its dimensions

56
6

Maria Luisa Sapino (BDMM 2010)

Cost of a dimension

- Every rectangle has to have values describing all its dimensions

Maria Luisa Sapino (BDMM 2010)

56
7

il nodo è associato a una pagina di memoria:
se la dimensione dell'informazione è molto grande (= ha molte dimensioni) occupo più spazio, quindi memorizzo meno elementi in una pagina

se riesco a contrarre l'informazione posso memorizzare più elementi nella stessa pagina: voglio che gli elementi siano discriminati da meno features possibile

TV-trees

- Hierarchical
 - Leaves: objects (documents)
 - Internal nodes: Minimum Bounding Regions
 - Higher fan-out at the root
 - Lower fan-out at the leaves (or lower levels)

Maria Luisa Sapino (BDMM 2010)

56
8

struttura gerarchica in cui
* le foglie sono gli oggetti da rappresentare
* i nodi interni rappresentano regioni che sono le MBR che contengono tutti gli elementi dei figli

- alla radice ho maggiore fanout perché uso meno features, quindi ho meno info -> più nodi nella stessa pagina
- verso le foglie ho fanout minore perché uso più features

Node structure in TV-trees

- In R-trees, every node is a hyper-rectangle
- In TV-trees, every node has
 - a center (in k-dimensions)
 - a radius (defined in n-dimensions)

Maria Luisa Sapino (BDMM 2010)

56
9

ogni regione è caratterizzata da CENTRO e RAGGIO

se ho m dimensioni, non voglio usarle tutte se utilizzarne di meno è sufficiente per discriminare gli oggetti:

* ordino le dimensioni in base al loro potere discriminante
* per ogni nodo suddivido le dimensioni in 3 gruppi:

- 1) quelle inutili
- 2) componenti attive rispetto alle quali si calcola la distanza dal centro, ovvero si definisce la regione -> rispetto a cui calcolo le distanze
- 3) quelle più discriminanti (le prime k), che assumono lo stesso valore per tutti gli elementi che stanno nella sottoregione rappresentata dal nodo -> non le uso nel calcolo delle distanze

un nodo è definito da:

- * coordinate del centro
- * raggio (scalare)

Node structure in TV-trees

- In R-trees, every node is a hyper-rectangle
- In TV-trees, every node has
 - a center (in k-dimensions)
 - a radius (defined in n-dimensions)

57
0

Maria Luisa Sapino (BDMM 2010)

le componenti del centro [1-k] si contraggono o espandono a seconda delle necessità,
es: se aggiungo un elemento rispetto a cui devo discriminare su una componente tra 1 e k devo CONTRARRE, perché quella feature diventa attiva (serve per discriminare!)

TV trees: example

57
1

•C, the center, has only one dimension, x
•Radius has only one dimension, y

Maria Luisa Sapino (BDMM 2010)

regione che ha $x = 2$ e y compresa tra -1 e 1 (non considero la componente z)

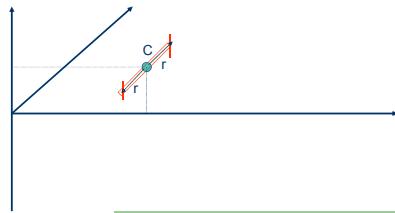
per calcolare le distanze che definiscono la regione uso solo la componente y!

TV trees: example

57
2

•C, the center, has only one dimension, x
•Radius has only one dimension, y
.....any z is okay

TV trees: extension example



57
3

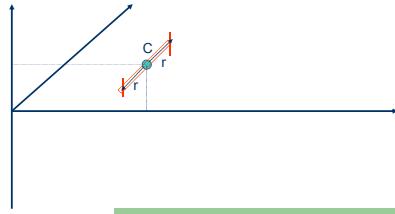
- C, the center, has only two dimensions, x,y
- Radius has only one dimension, z

Maria Luisa Sapino (BDMM 2010)

qui invece ho bisogno di più informazione per definire
una classe, quindi espando il vettore:

* k si sposta avanti di 1 -> uso x e y per calcolare il
centro (sono fissi), e la regione di variazione dei valori
ora è data dalla dimensione z

TV trees: extension example



57
4

- C, the center, has only two dimensions, x,y
- Radius has only one dimension, z
-any z is not okay!!!!!!

LIMITE: bisogna conoscere il potere discriminante
delle features per poterle ordinare

Drawback

- Information about the behaviour of single attributes, e.g., their selectivity, is required

57
7

Maria Luisa Sapino (BDMM 2010)

INSERIMENTO: scendo nell'albero scegliendo in quale regione inserire l'elemento, secondo i seguenti criteri:

- 1) privilegio minimo overlapping
- 2) privilegio regione in cui ho il maggior numero di dimensioni in comune e per cui quindi cresce meno il livello di dettaglio delle features da rappresentare

Dimensionality curse

- Exponential growth in the number of pointers needed, wasted storage, exponential subqueries (quadtrees)
- Larger MBRs means smaller fanout in trees and this is bad
- ...and...

58
5

Maria Luisa Sapino (BDMM 2010)

- se crescono le dimensioni devo memorizzare più informazioni per ogni regione e cresce lo spazio occupato dalla rappresentazione di ciascuna regione

- se crescono le MBR allora si ha un fanout minore dell'albero -> albero diventa più profondo, quindi i cammini dalla radice alle foglie crescono

struttura a indice è più efficiente se scendendo verso le foglie cresce la specializzazione

Dimensionality Curse

- Consider a query point and three alternative ranges:

58
6

Maria Luisa Sapino (BDMM 2010)

se in uno spazio 3D gli elementi sono distribuiti in modo uniforme, quando il range delle query cresce di un certo valore r il numero dei risultati cresce con la potenza delle dimensioni: quindi al crescere del range diventa una ricerca sequenziale!

Dimensionality Curse

- Consider a query point and three alternative ranges:

58
7

Maria Luisa Sapino (BDMM 2010)

se gli elementi sono distribuiti uniformemente il volume mi dice quanti elementi ci sono nella sfera: il rapporto tra le due sfere più esterne è molto superiore al rapporto tra quelle interne -> le query di range finiscono per restituire troppi elementi -> per questo motivo non si usano indici quando ci sono troppe dimensioni

Dimensionality Curse

- In n-dimensional space, if the number of points in the inner most sphere is I , then
 - number of points in the second slice is $O(2^{n-1} I)$
 - number of points in the second slice is $O(3^{n-1} I)$
 - number of points in the second slice is $O(4^{n-1} I)$
- This means that most of the points lie in the outermost slice!!!!

58
8

Maria Luisa Sapino (BDMM 2010)

Pyramid trees (Berchtold, Bohm, Kriegel, SIGMOD98)

- Motivation: drawbacks of already existing multidimensional index structures
 - Querying and indexing techniques which provide good results on
 - low-dimensional data do not perform sufficiently well on multi-dimensional data (curse of dimensionality)
 - high cost for insert/delete operations
 - Poor support for concurrency control/recovery

58
9

Maria Luisa Sapino (BDMM 2010)

I B-Tree, B⁺-Tree sono strutture ben note e con buone proprietà.
Siamo sicuri di non poterli utilizzare anche per indicizzare dati multidimensionali?
Torniamo a riconsiderare questa eventualità.

Idea: convertire lo spazio multidimensionale in uno monodimensionale in modo tale da associare un numero reale ad un dato così da poter utilizzare gli indici monodimensionali.

Pyramid tree

- Space-filling curves were using B-trees
- Pyramid trees also do the same..without space filling curves

59
0

Maria Luisa Sapino (BDMM 2010)

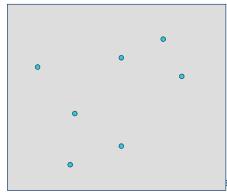
Pyramid Tree: suddivide lo spazio del database in piramidi nello spazio. Ad esempio nel caso tridimensionale, dato il parallelepipedo che identifica l'intero spazio, si disegnano 6 piramidi che partono ciascuna da una delle 6 facce del parallelepipedo ed hanno come vertice tutte il centro del parallelepipedo. Nel caso a 2 dimensioni le basi delle piramidi (piramidi 2D sono triangoli - quindi le basi sono lati dei triangoli) sono ai lati del rettangolo che denota lo spazio 2D e vanno verso il centro del rettangolo, ossia dello spazio.

SI ottengono 2^n piramidi dove n è il numero di dimensioni dello spazio.

Ogni piramide viene suddivisa in fette. Ciascuna fetta contiene dati mappati nello stesso numero reale.

Pyramid tree

- Space-filling curves were using B-trees
- Pyramid trees also do the same..without space filling curves



aria Luisa Sapino (BDMM 2010)

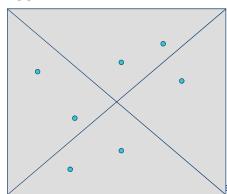
59
1

il modo in cui viene fatto il mapping è diverso da quello delle space-filling curves

* abbiamo spazio bidimensionale in cui si collocano i punti

Pyramid tree

- Space-filling curves were using B-trees
- Pyramid trees also do the same..without space filling curves



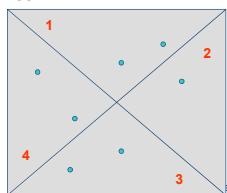
aria Luisa Sapino (BDMM 2010)

59
2

* pyramid tree usa l'elemento centrale dello spazio (punto di intersezione delle diagonali) come vertice di 2^D piramidi, dove D è il numero di dimensioni -> nel caso di spazio bidimensionale = 4 piramidi

Pyramid tree

- Space-filling curves were using B-trees
- Pyramid trees also do the same..without space filling curves



aria Luisa Sapino (BDMM 2010)

59
3

le piramidi vengono numerate in qualche modo, es. semplicemente numerandole in senso orario

Come caratterizzare le varie piramidi

indice < 2 coordinata di interesse (indice) a 0
indice ≥ 2 coordinata di interesse (indice mod 2) a 1

In generale ogni piramide di indice i ha base in

$d_indice = indice \bmod D$

ciascun punto (p_x, p_y) finisce in una piramide: come capire in quale piramide finisce?

Pyramid tree

- Space-filling curves were using B-trees
- Pyramid trees also do the same..without space filling curves

59
4

Considero la distanza tra il punto ed il punto centrale (c_x, c_y).

1) Se $|p_x - c_x| > |p_y - c_y|$
coordinata pari

- a) se $p_x < c_x$ piramide 4
b) se $p_x > c_x$ piramide 2

2) Se $|p_y - c_y| > |p_x - c_x|$
coordinata dispari

- a) se $p_y < c_y$ piramide 3
b) se $p_y > c_y$ piramide 1

Si "affetta" la piramide in tanti strati in cui ciascuno strato è identificato dalla sua distanza dal centro. Ogni strato ad esempio nel caso 2D, è costruito tracciando un segmento (o piano nel caso 3D) parallelo alla base.

Pyramid tree

- Space-filling curves were using B-trees
- Pyramid trees also do the same..without space filling curves

59
5

Ogni fetta è identificata da una coppia

<id piramide, id fetta cioè sua distanza dal centro>

I valori indicizzati dai B-Tree sono

<id piramide, id fetta> U punti che si trovano in quella regione

Le fette dovranno essere suddivise in modo tale da riservare una sola pagina per fetta.

Ottimo per query di tipo top-k: perché basta cercare una fetta

Svantaggioso per query di tipo range: occorre ricercare tutte le fette in cui possono esserci dati sufficientemente vicini.

se i dati sono distribuiti uniformemente, le fette più lontane dal vertice sono più soggette ad avere spazi vuoti

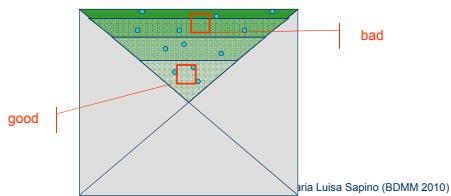
Pyramid tree

- If data is uniformly distributed, pages are likely to be of the same volume

59
6

Pyramid tree

- If data is uniformly distributed, queries likely to avoid thin pages, reducing the average access time



59
7

Maria Luisa Sapino (BDMM 2010)

problema: trovare una certa stringa in un certo testo

perché ci serve questo tipo di ricerca? --> perché l'indicizzazione del testo avviene tramite LISTE INVERTITE, ovvero indici che funzionano in modo inverso rispetto agli indici tradizionali dei libri:

- * nell'indice di un libro trovo info sulle parole chiave del libro associate alle pagine in cui trovarle
- * invece con le liste invertite, date una parola chiave mi dice in quali libri la trovo!

per ogni parola chiave, vogliamo rimandare ai documenti che la contengono -> dobbiamo trovare le parole chiave in un bacino di documenti

32
5

Keyword Search

- Full text scanning...

Maria Luisa Sapino (BDMM 2010)

Indicizzazione del testo

per rispondere a query del tipo

trovare tutti i documenti che parlano di un determinato argomento

32
6

Keyword Search

- Inversion
 - For each document maintain a list of matching documents
 - B-tree
 - Hashing
 - trie

Maria Luisa Sapino (BDMM 2010)

L'idea è quella di utilizzare gli indici invertiti, strutture dati in cui si mantiene l'informazione inversa nel senso che invece di associare un documento alle sue parole chiave, associo a ogni chiave una lista di documenti.

32
7

Trie (prefix)...

Maria Luisa Sapino (BDMM 2010)

Ogni nodo interno contiene l'alfabeto (A,B,C,...).

Radice: tutti i documenti

figlio della radice associato alla lettera A: tutti i documenti che contengono una keyword che inizia con la lettera 'A'. Ancora il figlio di tale nodo associato ad 'A' conterrà tutti i documenti che contengono parole chiave che iniziano con 'AA'.

Ogni foglia è un insieme di documenti che contengono la parola chiave identificate dal nodo (es. cammino CAT porta a documenti che contengono la parola chiave CAT).

Vantaggi: semplicità d'uso e di costruzione e operazioni veloci (es. ricerca lineare nella lunghezza della stringa da cercare).

Funzionano bene per la ricerca di prefissi (es. trovare tutti i documenti con keyword che iniziano con CAS).

Svantaggi: funziona solo per query esatte e bisogna conoscere tutti i termini su cui l'utente può fare ricerca -> parole esterne a tale vocabolario non verranno mai associate a nessun documento.

Non funziona con la ricerca di suffissi (es. parole che finiscono con SA)

se ho completato un termine del vocabolario, nel pallino verde avrò anche la lista dei documenti che contengono quella parola

32

Suffix Trees and Arrays

- Tries work well if the data consists of keywords...
- What if we do not have keywords?

Maria Luisa Sapino (BDMM 2010)

32

Suffix Trees and Arrays

- Tries work well if the data consists of keywords...
- What if we do not have keywords?
- Suffix trees and suffix arrays
 - Input text: a single long string
 - each position in the text gives a suffix

K. Selcuk Candan is teaching suffix trees in CSE515
Maria Luisa Sapino (BDMM 2010)

33

Suffix Trees and Arrays

- Suffix trees and suffix arrays
 - Input text: a single long string
 - each position in the text gives a suffix
- Text of length of N gives N suffixes

K. Selcuk Candan is teaching suffix trees in CSE515
Maria Luisa Sapino (BDMM 2010)

Per la ricerca basata sui suffissi viene fatta con Suffix Tree e Array

le ricerche possono però essere per SUFFISSO: "trova tutte le stringhe che terminano con una certa sottostringa"

* suffix tree
* suffix array (= funziona come uno heap -> uso indici dell'array per fare accesso più veloce)

DUE POSSIBILI IDEE:

1) ciascun carattere di una stringa identifica un suffisso, che va da quel carattere alla fine della stringa

se ho N caratteri ho N suffissi diversi

Suffix Trees and Arrays

Suffix Trees and Arrays

- Suffix trees and suffix arrays
 - Input text: a single long string
 - each position in the text gives a suffix
 - Text of length of N gives N suffixes
 - ..alternatively, text with W words give W suffixes,

Maria Luisa Sapino (BDMM 2010)



2) posso cercare solo sequenze significative, cioè
parole complete -> posso definire i suffissi usando solo
gli indici associati alle parole

se ho W parole ho W suffissi diversi

Suffix Trees

- Suffix trees

- Input text: a single long string
- each word position in the text gives a suffix

K. Selcuk Candan is teaching suffix trees in CSE515

The diagram shows 7 suffixes of the string "K. Selcuk Candan" starting from index 1:

- 1: K.
- 2: .Selcuk
- 3: .Selcuk C
- 4: .Selcuk Can
- 5: .Selcuk Can d
- 6: .Selcuk Can da
- 7: .Selcuk Can dan

A suffix tree for the string "K. Selcuk Candan". The root node is orange. Edges are labeled with characters: 'a' (from root to node 3), 'c' (from root to node 1), 'k' (from node 1 to node 2), 's' (from node 2 to node 3), 't' (from root to node 4), 'r' (from node 4 to node 5), 'e' (from node 5 to node 6), 'u' (from node 6 to node 7), and 'z' (from node 7 to node 3).

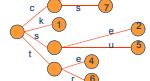
Maria Luisa Sapino (BDMM 2010)

33
2

Suffix Trees

- Suffix trees
 - Input text: a single long string
 - each word position in the text gives a suffix

K. Selcuk Candan is teaching suffix trees in CSE515



Maria Luisa Sapino (BDMM 2010)

* associamo un numero ad ogni parola in modo ordinato
 $W = 7$

ho 7 suffissi che indicizzo in questo modo:
* la radice ha tanti figli quanti sono i diversi caratteri con cui iniziano i termini da indicizzare, ordinati alfabeticamente

* da ciascun nodo di livello 1, continuo differenziando in base a quali parole iniziano con quella lettera,
nell'es: per C ho due alternative:

- A (CAN)
- S (CES515)

Suffix Trees

- Suffix trees

- Input text: a single long string
- each word position in the text gives a suffix

K. Selcuk Candan is teaching suffix trees in CSE510

The diagram shows seven suffixes of the string "K. Selcuk Candan" starting from index 1:

- 1: K.
- 2: .Selcuk
- 3: .Selcuk C
- 4: .Selcuk Ca
- 5: .Selcuk Can
- 6: .Selcuk Cand
- 7: .Selcuk Cand

Patricia trie is a trie where
all unary paths are compressed

A Patricia trie structure for the string "K. Selcuk Candan". The root node has children 'K' and 'C'. The 'K' child leads to a leaf node 'a'. The 'C' child leads to a node with children 'a' and 'n'. The 'a' child leads to a leaf node 'd'. The 'n' child leads to a node with children 'a' and 'd'. The 'a' child leads to a leaf node 'n'. The 'd' child leads to a leaf node 'a'. The 'n' child leads to a node with children 'e' and 's'. The 'e' child leads to a leaf node 'n'. The 's' child leads to a leaf node 'a'. The 'e' child leads to a leaf node 'a'. The 's' child leads to a node with children 't' and 'e'. The 't' child leads to a leaf node 'n'. The 'e' child leads to a leaf node 'a'. The 'e' child leads to a leaf node 'a'. The 's' child leads to a node with children 't' and 'e'. The 't' child leads to a leaf node 'n'. The 'e' child leads to a leaf node 'a'. The 'e' child leads to a leaf node 'a'.

This is also a Patricia trie

Maria Luisa Sapino (BDMM 2010)

33

3

Suffix Trees

- Suffix trees
 - Input text: a single long string
 - each word position in the text gives a suffix

A horizontal timeline consisting of seven red rectangular boxes labeled 1 through 7. Each box contains a small icon of a person wearing a graduation cap and holding a book, representing a teacher.



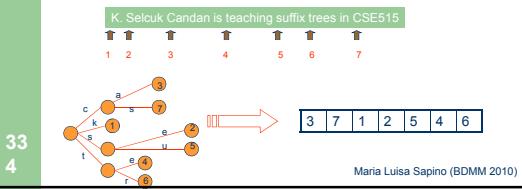
le due parole sono sufficienti a discriminare
completamente i termini che iniziavano con C: quindi ho
compattato in un unico nodo quello che sarebbe stato
un cammino lineare per ciascuna parola, ovvero una
lista

=
un albero in cui i cammini lineari sono compatti in
un'unica foglia si chiama Patricia trie

* il numero della foglia rappresenta l'indice che avevo
associato alla parola nella stringa

Suffix Arrays

- Suffix trees
 - Input text: a single long string
 - each word position in the text gives a suffix

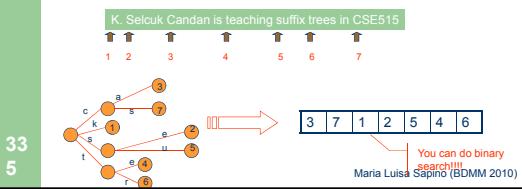


perché è importante inserire le lettere in modo ordinato?

per passare alla rappresentazione ad array della stessa informazione, semplicemente leggendo in ordine le foglie -> le posizioni rappresentano i suffissi individuati in ordine alfabetico

Suffix Arrays

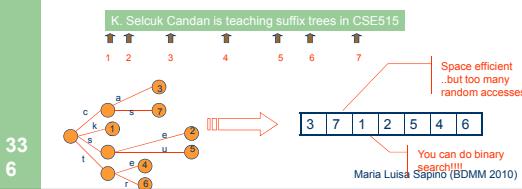
- Suffix trees
 - Input text: a single long string
 - each word position in the text gives a suffix



vantaggio: posso fare ricerca binaria (perché sono ordinati) che ha costo $\log(n)$

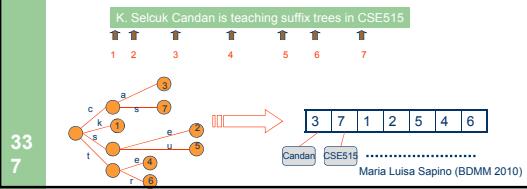
Suffix Arrays

- Suffix trees
 - Input text: a single long string
 - each word position in the text gives a suffix



Suffix Arrays

- Suffix trees
 - Input text: a single long string
 - each word position in the text gives a suffix



Nei suffix tree i figli di ogni nodo vengono ordinati in ordine alfabetico (es. dalla radice dell'albero nel lucido, i figli sono ordinati in modo tale che le lettere degli archi corrispondenti siano in ordine alfabetico -> t,s,k,c).

In questo modo, una visita dell'albero in ordine permette di enumerare i suffissi in ordine alfabetico (è possibile memorizzare tale lista in un array detto **suffix array**).

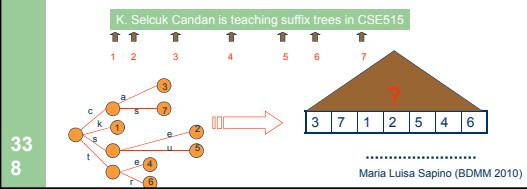
Vantaggi del suffix tree

- 1) posso fare ricerca binaria (quindi in $\log(n)$)

- 2) posso costruire un altro indice, es. un B-Tree

Suffix Arrays

- Suffix trees
 - Input text: a single long string
 - each word position in the text gives a suffix



non abbiamo a priori la possibilità di indicizzare i documenti in cui fare la ricerca

...no arrays..no suffixes??

- Can we do search without a data structure on text?

33
9

Maria Luisa Sapino (BDMM 2010)

...no arrays..no suffixes??

- Can we do search without a data structure on text?
- Create a data structure on the query!!!!

34
0

Maria Luisa Sapino (BDMM 2010)

creiamo una struttura di dati sulla query!

supponiamo di avere a disposizione tutti i libri di fiabe e di cercare la parola "abracadabra": voglio restituire tutte le occorrenze della parola in una qualsiasi fiaba

(senza aver indicizzato le parole dei documenti o cmq non avendo quella parola nell'indice)

...no arrays..no suffixes??

- Given a text of length N....and pattern M
- Brute force: $O(NM)$

Average behavior closer to $O(N)$
-errors are found quick!!!

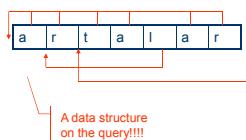
34
1

Maria Luisa Sapino (BDMM 2010)

Troppo costoso $O(NM)$!

...no arrays..no suffixes??

- Given a text of length N....and pattern M
- Knuth-Morris-Pratt: $O(N)$



34
2

Maria Luisa Sapino (BDMM 2010)

imporre una struttura per riconoscere il pattern oggetto della query:

si parte dalla stringa di cui cerco il pattern, es:

ARTALAR

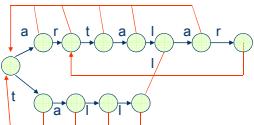
- 1) se riconosco il carattere passo al successivo
- 2) se arrivo ad un carattere che non riconosco (es: se ho R al posto della prima L) si può ricominciare la ricerca sapendo di aver già riconosciuto la A immediatamente precedente la L

bruteforce: ad ogni fallimento ricomincio da capo

ogni freccia mi dice da dove devo ricominciare se non trovo un match in quel punto

...no arrays..no suffixes??

- Given a text of length N....and **multiple patterns**
- Aho-Corasick Trie: $O(N)$



Maria Luisa Sapino (BDMM 2010)

34
3

...no arrays..no suffixes??

- Given a text of length N....and pattern M
- Boyer-Moore: $O(N(\log M)/M)$

Maria Luisa Sapino (BDMM 2010)

34
4

sfrutta l'informazione sulla struttura della query per contare il numero di posizioni di cui si deve arretrare la ricerca: anziché partire dal primo elemento della stringa data e dal primo elemento della stringa che si sta cercando, parte dall'ultimo elemento della stringa che sto cercando

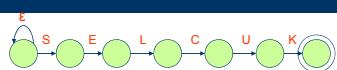
es. sto cercando stringa lunga 8:
controllo se l'elemento 8 coincide -> se non coincide so per certo che nelle prime 8 posizioni non c'era la stringa che cercavo

ora

- se il carattere in posizione 8 non appartiene alla stringa da cercare allora ricomincio dal carattere successivo
- altrimenti vado all'indietro e guardo se c'è la stringa che mi interessa

Nondeterministic Finite Automaton

NFA...upto 0 errors



Maria Luisa Sapino (BDMM 2010)

34
5

finora abbiamo visto solo query esatte: ora vogliamo introdurre una possibilità di errore, quindi ammettiamo di trovare stringhe "simili" con un certo errore.
Questo viene fatto perché spesso è ammissibile nella pratica avere fare matching con qualche errore, così da poter utilizzare algoritmi più efficienti.

NFA...upto 1 insertion

34
6

Maria Luisa Sapino (BDMM 2010)

consentono un errore di inserimento di al più un carattere nella stringa che sto cercando.

Si inizia dallo stato in alto a sinistra. Lo * denota

'qualsiasi carattere'. Se da uno stato si riconosce una lettera errata, si salta agli stati nella parte inferiore del lucido.

Esempio: se riconosco SEL sono nello stato 4 della fila in alto nel lucido. Se il carattere successivo non è C si passa alla fila in basso ma il carattere successivo deve essere C. Quindi una stringa matcha con SELCUK se è uguale oppure se ha un carattere in più e quindi se è ottenuta da SELCUCK con un'operazione di inserimento di un carattere. Per questo si dice che questo automa consente l'errore di

NFA...upto 1 insertion\replacement

34
7

Maria Luisa Sapino (BDMM 2010)

consentono sia errore di inserimento che di sostituzione di un carattere, ma solo uno dei due!

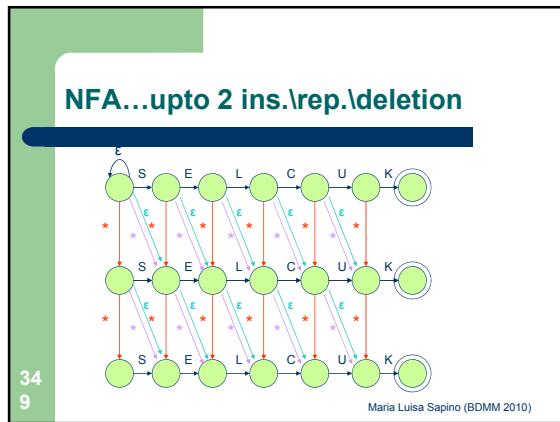
NFA...upto 1 ins.\rep.\deletion

34
8

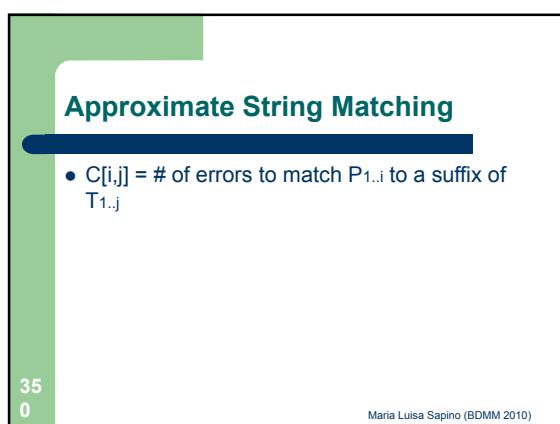
Maria Luisa Sapino (BDMM 2010)

consentono un errore di inserimento o cancellazione o sostituzione (solo uno dei tre!)

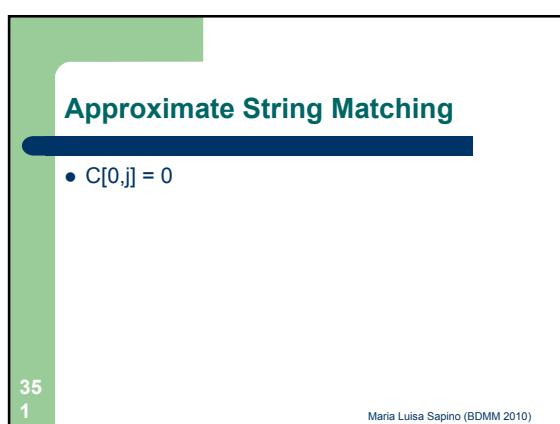
--> permette di riconoscere stringhe simili a quella che cercavo con al più un errore (di qualsiasi natura)



34
9



35
0



35
1

obiettivo: contare, data una stringa T e dato un pattern P che voglio riconoscere all'interno della stringa, il # di errori nel cercare di riconoscere il pattern come suffisso della stringa

-
- * i è la lunghezza del pattern
 - * j la posizione all'interno del suffisso
-

NB: mentre la costruzione di un automa (che può essere utilizzato anche per verificare la distanza tra due stringhe) va fatta per ogni stringa da confrontare, la distanza di edit permette di fare questi confronti con un unico algoritmo, semplice e compatto e generale (valido per ogni stringa, senza costruire nessun automa).

sto cercando, all'interno del suffisso in cui sono posizionato in posizione j , il pattern vuoto: siccome mi manca solo più lo 0-esimo elemento, lo trovo con costo = 0

35
2

Approximate String Matching

- $C[0,j] = 0$
- $C[i,0] = i$

Maria Luisa Sapino (BDMM 2010)

sto cercando il pattern da 1 a i in una stringa in cui sono posizionata sulla posizione 0 (= stringa vuota) : faccio i errori (= i inserimenti)

35
3

Approximate String Matching

- $C[0,j] = 0$
- $C[i,0] = i$
- if($P_i = T_j$)
 - $C[i,j] = C[i-1,j-1]$

Maria Luisa Sapino (BDMM 2010)

se l'ultimo elemento della stringa è uguale all'ultimo elemento del pattern che sto cercando, allora non faccio errori

35
4

Approximate String Matching

- $C[0,j] = 0$
- $C[i,0] = i$
- if($P_i = T_j$)
 - $C[i,j] = C[i-1,j-1]$
- else
 - $C[i,j] = 1 + \min\{C[i-1,j], C[i,j-1], C[i-1,j-1]\}$

Maria Luisa Sapino (BDMM 2010)

altrimenti c'è un errore:
 il numero di errori fino a quel punto è il minimo tra
 1) il # di errori per cercare la stringa fino all'elemento i-esimo in posizione j-1
 2) il # di errori per cercare la stringa fino all'elemento i-1-esimo in posizione j
 3) il # di errori per cercare la stringa fino all'elemento i-1-esimo in posizione j-1

+

1 errore dovuto a

- 1) inserimento
- 2) cancellazione
- 3) sostituzione

Approximate String Matching

- $C[0,j] = 0$
- $C[i,0] = i$
- if($P_i = T_j$)
 - $C[i,j] = C[i-1,j-1]$
- else
 - $C[i,j] = 1 + \min\{C[i-1,j], C[i,j-1], C[i-1,j-1]\}$

$O(MN)$

Maria Luisa Sapino (BDMM 2010)

35
5

tecnica da considerare per gestire dati multidimensionali di dimensionalità elevata: HASHING
una funzione hash mappa un oggetto in un numero che, con probabilità altissima, è diverso dai numeri associati agli altri oggetti -> una buona funzione di hashing minimizza il # di collisioni

5/26/2011

--> vogliamo usare le funzioni hash cercando di avere collisioni quando gli oggetti associati allo stesso numero sono molto simili

1

Hashing for nearest neighbor search

- Hashing generally works for "equality searches"
- ..can we use "hashes" for nearest-neighbor searches???
-if they are **locality sensitive**, then "yes"!

* l'hashing è definito per ricerche per uguaglianza: dato un oggetto, faccio l'hashing e spero di trovarlo nella posizione data dalla funzione

* ora noi vogliamo usare questa funzione per la ricerca NN: è possibile farlo quando le funzioni di hash sono **LOCALITY SENSITIVE**, ovvero il valore restituito dalla funzione di hashing è sensibile alla somiglianza tra gli oggetti che vengono mappati, ossia c'è probabilità molto elevata che gli oggetti simili vengano mappati nello stesso bucket

2

Locality Sensitive Hashing (LSH)

- What is "locality sensitive hashing"?
 - ...a "grid" is a locality sensitive hash
 - ...a space filling curve is a locality sensitive hash
 - More specifically, these are **deterministic** functions that tend to map nearby points to the same or nearby values.
- Can we develop **randomized** locality sensitive hashes?

stessa idea dei GRID FILES: oggetti vicini nello spazio finiscono nella stessa cella

anche SPACE FILLING CURVE: mappano dati k-dimensionali in dati monodimensionali che conservano le distanze

--> funzioni deterministiche che tendono a mappare punti vicini tra loro in valori che sono uguali o cmq molto vicini tra di loro

vogliamo estendere il concetto di hashing **deterministico a concetto di hashing randomizzato** (probabilistico) con la caratteristica di preservare la distanza tra gli oggetti

3

Locality Sensitive Hashing (LSH)

- Let $sim()$ be a similarity function
- A locality sensitive hash corresponding to $sim()$ is a function, $h()$, such that

$$prob(h(o1) = h(o2)) = sim(o1, o2)$$

- The challenge is to find the appropriate $h()$ for a given $sim()$

idealmente, data una funzione di somiglianza tra due oggetti, si vuole che la misura di somiglianza tra due oggetti sia coerente con i valori restituiti dalla funzione di hashing, ossia che la probabilità di avere una collisione tra $h(o1)$ e $h(o2)$ sia uguale alla loro somiglianza:
* se $o1 = o2$ allora i due oggetti sono identici, quindi $sim(o1, o2) = prob(h(o1) = h(o2)) = 1$
* altrimenti $sim(o1, o2) = prob(h(o1) = h(o2)) = P$

la probabilità che gli oggetti vengano mappati nello stesso valore di h è tanto maggiore quanto più i due oggetti sono simili

ovviamente poi dovremo mappare la query nello stesso modo per confrontarla con gli oggetti che abbiamo mappato con h

Locality Sensitive Hashing (LSH)

- An LSH family, H , is (r, cr, P_1, P_2) -sensitive, if for any two objects o_i and o_j and for a randomly selected $h \in H$
 - if $dist(o_i, o_j) \leq r$ then $prob(h(o_i) = h(o_j)) \geq P_1$,
 - if $dist(o_i, o_j) \geq cr$ then $prob(h(o_i) = h(o_j)) \leq P_2$ and
 - $P_1 > P_2$.
 - cr multiplo di r

Per trovare una funzione hash che abbia la proprietà appena descritta (nel lucido precedente) si procede nel modo seguente:

Si definisce una famiglia di funzioni tramite una quadrupla

r: raggio inferiore
cr: raggio superiore
P1: probabilità che due oggetti simili collidano
P2 : probabilità che oggetti diversi non collidano

La famiglia di funzioni (r, cr, P_1, P_2) contiene tutte e sole le funzioni tali che

- se $dist(o_1, o_2) \leq r$, ossia se o_1 e o_2 sono abbastanza simili (e lo sono se la distanza è entro il raggio inferiore 'r') allora $prob(h(o_1)=h(o_2)) \geq P_1$, ossia con probabilità almeno P_1 i due oggetti collidono
- se $dist(o_1, o_2) \geq cr$ ossia se o_1 e o_2 sono abbastanza distanti allora $prob(h(o_1)=h(o_2)) \leq P_2$, ossia con probabilità al più P_2 i due oggetti non collidono
- cr è un multiplo di r
- $P_2 < P_1$ perché la probabilità che due oggetti collidano quando sono simili deve essere (molto) maggiore della probabilità che due oggetti diversi collidano.

Locality Sensitive Hashing (LSH)

- Consider a (r, cr, P_1, P_2) -sensitive hash family, H
 - Let's create L composite hash functions
- $$g(o) = (h_{1,1}(o), \dots, h_{k,L}(o))$$
- by picking $L \times k$ hash functions, $h_{i,j} \in H$, independently and uniformly at random from H .

Sul dominio degli oggetti del database, si costruiscono funzioni di hash composte (da altri funzioni hash).

L'idea è quella di applicare più funzioni hash (prese dalla famiglia appena descritta) ad una coppia di oggetti. Se gli oggetti sono simili, la probabilità che una funzione hash li faccia collidere è elevata (ma non è certo). Però applicando più funzioni hash a tale coppia, il fatto che la maggior parte di esse le faccia collidere rafforza il fatto che i due oggetti debbano collidere (più funzioni si applicano più è bassa la probabilità che la maggior parte di tali funzioni facciano collidere i due oggetti erroneamente).

Locality Sensitive Hashing (LSH)

- Let us be given $g_1()$ through $g_L()$ and database, D ,
 - Hash object o in D using $g_1()$ through $g_L()$ and include o in all matching hash buckets
- $$g_1(o) = (h_{1,1}(o), \dots, h_{k,1}(o)),$$
-
- $$g_L(o) = (h_{1,L}(o), \dots, h_{k,L}(o))$$

quando abbiamo l'oggetto da inserire nel db, in quale bucket lo inseriamo?

--> lo inseriamo in tutti gli L bucket in cui l'oggetto viene mappato dalla funzione di hashing

7

Locality Sensitive Hashing (LSH)

- Hash the query q in also using $g_1()$ through $g_L()$ and consider all objects in these hash buckets

$$g_L(q) = (h_{1,L}(q), \dots, h_{k,L}(q)),$$

$$\dots$$

$$g_L(q) = (h_{1,L}(q), \dots, h_{k,L}(q))$$
- Key result:
 • if $L = \log_{1-\rho} \delta$, then any object within range r is returned with probability at least $1-\delta$.

quando ho una query, mappo anche la query usando le L funzioni composite e considero come oggetti simili quelli che erano stati indicizzati nei bucket a cui la query stessa è stata mappata

RISULTATO alla base dell'approccio:

se il numero di bucket, L, è = $\log_{1-P_1} \delta$ (\delta), allora ogni oggetto che dista al più r dalla query viene restituito con probabilità almeno pari a $1 - \delta$.

Quindi giocando sul valore di L ho delle garanzie sulle probabilità di restituzione dei risultati corretti

8

Locality Sensitive Hashing (LSH)

- Then, how do we create a (r, cr, P_1, P_2) -sensitive hash family, H ??
-depends on the underlying $sim()$ or $\delta()$ function...

come costruire la famiglia di funzioni H?
dipende dalla funzione di somiglianza (o di distanza)

9

Locality Sensitive Hashing (LSH)

- Assume d-dimensional binary vector; e.g. $(0,1,1,1,0,\dots,1)$
- Let $\delta()$ be the hamming distance (number of differing dimensions between two vectors)
- H contains all projections of the input point x on one of the coordinates; i.e., $h_i(x) = x_i$

esempio di dominio:

- I dati del database sono vettori binari a n dimensioni (ogni dimensione può assumere i valori 0 o 1).
 - Sia la funzione di distanza δ la DISTANZA DI HAMMING (numero di bit che differiscono in due stringhe di bit, nel nostro caso numerod i dimensioni per cui differiscono due vettori).
 - La funzione hash $h_i(x_1, \dots, x_n) = x_i$, ossia la funzione che restituisce una dimensione
-
-
-

10

Locality Sensitive Hashing (LSH)

- Let
 - p and q be two vectors in d -dimensional binary vector space
 - $\delta()$ is the hamming distance
 - H contains $h_j(x) = x_j$
- Note that $\text{prob}[h(q) = h(p)]$ is equal to the fraction of coordinates on which p and q agree.
- Then, if we select
 - $P_1 = 1 - (r/d)$ and $P_2 = 1 - c(r/d)$
 - such that $c > 1$
 we have $P_1 > P_2$.

dati due vettori p e q in questo spazio, quale è la probabilità che $h(p)$ e $h(q)$ collidano?

--> la P di collisione è uguale alla percentuale di coordinate su cui c'è accordo

* la P minima di collisione tra oggetti molto vicini è
 $P_1 = 1 - r/d$

dove d è il numero di elementi in ciascun vettore e r è la somiglianza che voglio garantire: r/d è la percentuale di elementi che devono essere in accordo

* $P_2 = 1 - c(r/d)$, con $c > 1$

--> sicuramente $P_1 > P_2$

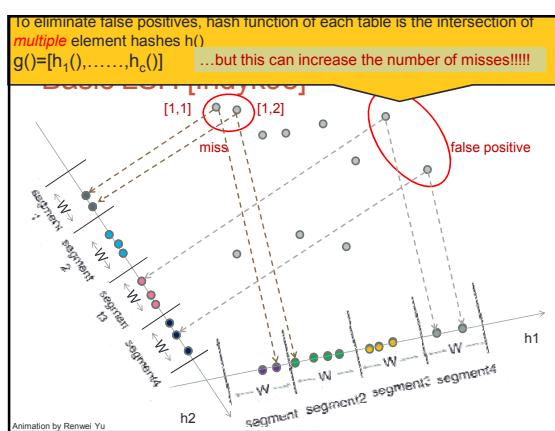
11

Locality Sensitive Hashing (LSH)

- L1-distance in d-dimensional space:**
 - pick a $w \gg r$
 - impose a randomly shifted grid with cells of width w
 - pick random s_1, s_2, \dots, s_d in $[0, w]$
 - define $h_{s_1, s_2, \dots, s_d}(x) = (|(x_1 - s_1)/w|, \dots, |(x_d - s_d)/w|)$.

Costruire una griglia nel seguente modo:

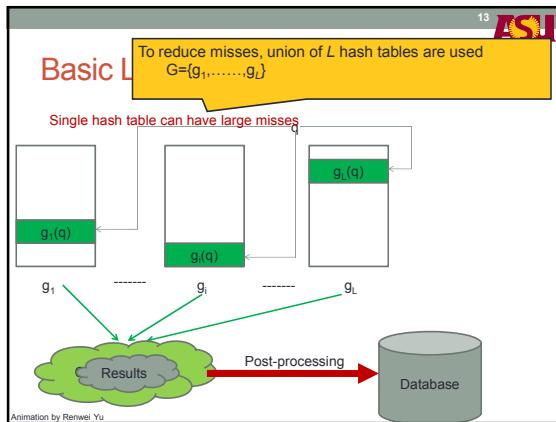
Si individuan d punti diversi per uno spazio a d dimensioni



proietto tutti i punti su una retta su cui definisco dei segmenti lunghi w : oggetti simili vengono mappati nello stesso segmento

* i falsi positivi erano vicini solo rispetto a h_1 , mentre con h_2 finiscono in due bucket diversi: la quantità di direzioni in cui i due punti vengono mappati insieme supporta la tesi della loro vicinanza

--> per eliminare i falsi positivi posso considerare come appartenenti allo stesso bucket solo gli oggetti che stanno nell'intersezione di diverse funzioni di hashing: questo ha dei rischi perché può causare dei MISS se due punti sono a ridosso del confine tra due bucket!



per evitare misses, si può fare UNIONE (invece che INTERSEZIONE) dei bucket ed eliminare i false hits in fase di post-processing

* prendo le singole tabelle e le considero come generatori di candidati su cui poi fare post-processing

Locality Sensitive Hashing (LSH)

- **Ls-distance in d-dimensional space:**
 - pick a $w \gg r$
 - pick a random projection, p , of the space onto a 1-dimensional line by picking each coordinate of p from the Gaussian distribution.
 - chop the line into segments of length w , shifted by a random value b in $[0, w]$; i.e., given vector x

$$h_{x,p}(x) = \lfloor (p \cdot x + b)/w \rfloor,$$

si può estendere questo approccio alle dimensioni della distanza Ls in cui:

* definisco finestra di dimensione $w \gg r$
* per ogni regione dello spazio scelgo una proiezione p dallo spazio di dimensione d ad una linea monodimensionale, scegliendo in modo casuale ciascuna coordinata di p secondo la distribuzione di probabilità Gaussiana
* segmento la linea in segmenti di dimensione w shiftati di un valore b scelto a caso da $[0, w]$
----> qui si fa hashing solo su una componente p

Issues of Basic LSH

- Large number of tables to achieve good search quality
 - $L > 580$ in [Buhler01]
- Impractical for large datasets, need reduce hash tables
 - Entropy-based LSH [Panigrahy06]
 - Multi-Probe LSH [Qin07]

this slide by Renwei Yu

PROBLEMI DI LSH:

1) per dimensioni di db molto grandi serve L molto grande

Issues of Basic LSH (continued...)

- Data dependent parameters need hand-tuning
- Different bucket size is required to collect enough candidates to answer different KNN queries
- LSH-Forest [Bawa05]
- Multi-Probe LSH can also be self-tuning to answer different KNN queries

this slide by Renwei Yu

2) r è un parametro alla base della definizione della famiglia di hashing, quindi la famiglia di funzioni funziona correttamente per query fatte su quel range r
---> in questo caso correliamo la probabilità di successo al range scelto inizialmente, quindi se si fanno query di range molto più piccolo si ha maggiore probabilità di false hits

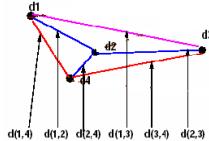
abbiamo visto diverse tecniche di indicizzazione tutte basate sul presupposto che noi conosciamo le features degli oggetti che indicizzavamo. cosa possiamo fare se non conosciamo le features?

è possibile che noi abbiamo informazioni sulla somiglianza reciproca tra due oggetti senza sapere come è stata misurata questa somiglianza, non ho la descrizione delle singole features: black-box

--> se non ho a disposizione delle features, ho dei modi per suddividere in cluster di oggetti somiglianti i miei dati

SE NON CONOSCO LE FEATURES POSSO USARE IL CLUSTERING, MA POSSO FARLO ANCHE SE CONOSCO LE FEATURES (a maggior ragione, visto che con le features posso calcolare la somiglianza!)

What if we do not have features??



We know the distances, but

- we do not have explicit features
- distances are not metric....

Maria Luisa Sapino (BDMM 2010)

60
3

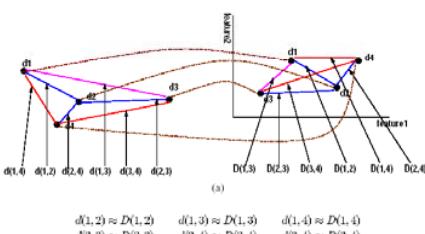
posso avere strumenti che danno una misura di somiglianza tra due oggetti senza dare motivazioni in base a proprietà oggettive, es: somiglianza di pagine web in base al fatto che sono spesso visitate da un utente in sequenza

--> dobbiamo creare schemi per suddividere i dati non in base alle features degli oggetti da indicizzare

posso anche provare a "indovinare" le features che non ho in modo tale da ottenere la distanza che conosco, però è comodo solo se devo combinare due oggetti, altrimenti diventa troppo complicato!

--> MULTIDIMENSIONAL SCALING

Multi Dimensional Scaling



60
4

si applica quando le distanze sono di tipo non-metrico (e quindi non posso applicare indici, perché funzionano solo con misure metriche per poter potare!)

allora vogliamo ricondursi ad uno spazio nuovo in cui possa definire una distanza METRICA su cui poter costruire un indice: nell'es. conosciamo tutte le distanze a due a due tra 4 punti, ma non conosciamo le singole features

MDS: si parte da un numero di features basso (es. 2) e si cerca una distribuzione dei punti nello spazio che sia tale per cui la distanza reciproca tra i punti sia il più possibile fedele a quella che conosco già -> cerco di determinare le coordinate dei punti

MDS

- The criterion for the mapping is to minimize stress

$$\text{stress} = \sqrt{\frac{\sum_{i,j} (d_{ij}^s - d_{ij}^c)^2}{\sum_{i,j} d_{ij}^2}}$$

- Start with a (random) configuration of points with low dimensions
- Apply some form of steepest descent iteratively to minimize the stress.
 - move objects
 - add dimensions

60
5

Maria Luisa Sapino (BDMM 2010)

in questo modo abbiamo le features e quindi abbiamo dei surrogati che possiamo indicizzare

vogliamo che la disposizione dei punti soddisfi al meglio la distanza data: per fare questo bisogna minimizzare una funzione di STRESS che calcola l'errore che io faccio nel confrontare le coppie di oggetti se, anziché utilizzare le distanze vere, io utilizzo le distanze che ho stimato (questo per ogni possibile coppia di punti) -> percentuale di errore, quindi va minimizzata

::: se le distanze fossero uguali allora il numeratore sarebbe = 0

come fare?

* si può generare una configurazione iniziale random per poche dimensioni

* poi si usa un procedimento iterativo di minimizzazione più ripido possibile, spostando gli elementi in modo da ridurre lo stress: quando la discesa rallenta molto, quindi non porta più a miglioramenti significativi, si prova ad aggiungere dimensioni

N.B. algoritmo è MONOTONO -> lo stress diminuisce ad ogni passo

MDS

- How to map the query?

60
6

Maria Luisa Sapino (BDMM 2010)

MDS è un processo costoso, e in più non si può fare una volta sola all'inizio: anche se costruisco lo spazio con tutti i punti mappati, dove metto la query nel mio spazio costruito artificialmente?

--> ad ogni query dovrei applicare lo stesso procedimento!!!

-) questa è la maggiore debolezza dell'MDS

N.B. se conoscessi già la distanza della query dagli oggetti non avrei nemmeno bisogno di farla!

Use of clusters (prune search space)

60
7

Maria Luisa Sapino (BDMM 2010)

SOLUZIONE: utilizzare meccanismi di clustering

CLUSTERING: identifico insiemi di oggetti che sono tra di loro a due a due simili e poi, per evitare ricerche sequenziali sugli oggetti (visto che non ho indici), uso ciascuna classe come gruppo ben rappresentato da un suo elemento (RAPPRESENTANTE del cluster) con cui effettuo i confronti

Use of clusters (prune search space)

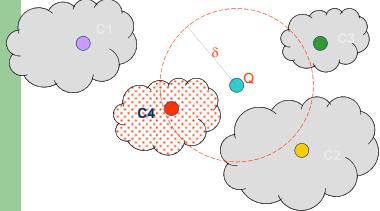
- ...given a query

60
8

Maria Luisa Sapino (BDMM 2010)

siccome si usano i rappresentanti per fare i confronti, è importante mettere nello stesso cluster oggetti simili

Use of clusters (prune search space)



- 60 • ...eliminate clusters based on their representatives
9

Maria Luisa Sapino (BDMM 2010)

quando si fa una query di range δ , si confronta Q con i rappresentanti dei cluster e si scartano i cluster i cui rappresentanti distano da Q più di δ (senza guardare gli altri elementi del cluster!)

Clustering methods

- Sound methods:
 - need a fixed document-to-document similarity matrix
- Iterative methods:
 - use document vectors iteratively

61
0

Maria Luisa Sapino (BDMM 2010)

DUE CLASSI DI METODI DI CLUSTERING:

- * metodi SOUND: portano ad un risultato esatto
---> si possono applicare quando conosco a priori tutti gli oggetti da clusterizzare e le loro distanze reciproche
 - * metodi ITERATIVI: portano ad un risultato approssimato
---> posso applicarli anche quando mi arriva un oggetto alla volta e lo devo aggiungere -> non devo conoscerli tutti fin da subito!
-
-
-
-

Outline of sound methods

- Find the similarity of each object pair

61
1

Maria Luisa Sapino (BDMM 2010)

Outline of sound methods

- Find the similarity of each object pair
- Setup a threshold
 - $\text{sim}(o_1, o_2) < T$ (objects are very different!)
 - $\text{sim}(o_1, o_2) \geq T$ (objects are comparable)

61
2

Maria Luisa Sapino (BDMM 2010)

Metodi sound:

Sceglio una soglia T di somiglianza. Ogni cluster deve contenere elementi simili almeno T a due a due, ossia A è simile almeno T da qualsiasi oggetto nel cluster.

Creo un grafo che rappresenta la somiglianza, in modo tale che

Outline of sound methods

- Find the similarity of each object pair
- Setup a threshold
 - $\text{sim}(o_1, o_2) < T$ (objects are very different!)
 - $\text{sim}(o_1, o_2) \geq T$ (objects are comparable)
- Create a graph which represents object similarities
 - Each pair of objects that are comparable is connected with an edge

61
3

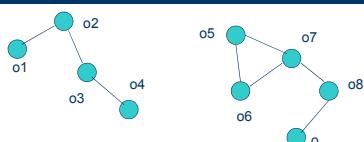
Maria Luisa Sapino (BDMM 2010)

1) misuro la somiglianza

2) definisco una soglia T che definisce la qualità del clustering: due oggetti che finiscono nello stesso cluster devono avere somiglianza $\geq T$

3) creo un grafo che rappresenti le somiglianze tra gli oggetti: grafo in cui ciascun vertice è associato ad un oggetto del db e due vertici sono connessi se e solo se la loro somiglianza è $\geq T$

Example collection



61
4

Maria Luisa Sapino (BDMM 2010)

le componenti connesse sono cluster

Connected components

61
5

Maria Luisa Sapino (BDMM 2010)

Connected components

- ...are o1 and o4 really similar?????

61
6

Maria Luisa Sapino (BDMM 2010)

problema: la somiglianza non è transitiva!

nell'es. o1 e o4 non sono simili, altrimenti sarebbero connessi -> questa strategia mi garantisce solo che se due oggetti sono in due cluster diversi allora sono diversi (hanno somiglianza < T)

Clique...

61
7

Maria Luisa Sapino (BDMM 2010)

dovrei individuare le componenti fortemente connesse, ovvero quelle in cui ogni vertice è connesso a tutti gli altri vertici del cluster

Clique...

- Clusters are overlapping!
- Costlier to compute (NP-complete)

Maria Luisa Sapino (BDMM 2010)

SVANTAGGI:

-) calcolo delle componenti fortemente connesse è NP-completo
 -) overlapping tra clusters nei punti di raccordo tra le componenti fortemente connesse -> questo significa dover considerare più gruppi per rispondere a una query
-
-
-
-
-

...or single-pass iterative method

- choose an object, and make it a cluster

Maria Luisa Sapino (BDMM 2010)

...or single-pass iterative method

- choose an object, and make it a cluster
- choose another object, \circ
 - find the closest cluster, c
 - if $dist(\circ, c) < T$ add \circ to c
 - else \circ is a new cluster

Maria Luisa Sapino (BDMM 2010)

...or single-pass iterative method

- choose an object, and make it a cluster
- choose another object, \circ
 - find the closest cluster, c
 - if $\text{dist}(\circ, c) < T$ add \circ to c
 - else \circ is a new cluster
- repeat until all objects are processed

62
1

Maria Luisa Sapino (BDMM 2010)

1) prendo un oggetto e lo considero come rappresentante di un nuovo cluster

2) per ogni oggetto che considero

- trovo il cluster più vicino

- se il cluster è vicino $< T$ aggiungi l'oggetto al cluster, altrimenti \circ è il rappresentante di un nuovo cluster

3) ripeti il passo 2 fino a che tutti gli oggetti non siano stati considerati

...or single-pass iterative method

- choose an object, and make it a cluster
- choose another object, \circ
 - find the closest cluster, c
 - if $\text{dist}(\circ, c) < T$ add \circ to c
 - else \circ is a new cluster
- repeat until all objects are processed

62
2

also called the "leader" algorithm

Maria Luisa Sapino (BDMM 2010)

algoritmo del LEADER:

ad ogni passo l'elemento che inserisco determina se verrà creato un nuovo cluster di cui quell'elemento sarà il leader

....how do we compute distance?



62
3

Maria Luisa Sapino (BDMM 2010)

....how do we compute distance?

62 • use a cluster representative...
4

Maria Luisa Sapino (BDMM 2010)

....how do we compute distance?

62 • use a cluster representative...
5

Maria Luisa Sapino (BDMM 2010)

....how do we compute distance?

62 • use a cluster representative...
6

Maria Luisa Sapino (BDMM 2010)

come calcolare le distanze?

-
- 1) uso il rappresentante, quindi la distanza di un oggetto da cluster è la distanza tra l'oggetto e il rappresentante del cluster
-
-
-
-

posso usare

-
- * rappresentante FISSO, ad es. il primo elemento che inserisco nel cluster
-

-
- +) è facile, perché se aggiungo un elemento al cluster non devo ricalcolare il leader
-

-
-) dopo qualche inserimento il leader potrebbe non essere più molto rappresentativo
-
-

-
- * rappresentante ADATTIVO, ricalcolato dinamicamente ad es. scegliendo sempre il punto che sta a distanza media da tutti gli altri elementi del cluster
-

-
- +) l'oggetto con cui rappresento il cluster è più rappresentativo, quindi i confronti funzioneranno meglio
-

-
-) è necessario ricalcolarlo ogni volta che si inserisce un nuovo elemento nel cluster
-
-

....alternatives...

Maria Luisa Sapino (BDMM 2010)

62
7

devo inserire un elemento con threshold T: in questo caso posso scegliere tra 3 cluster -> quale scelgo?

....alternatives...

Maria Luisa Sapino (BDMM 2010)

62
8

1) scelgo il primo cluster che trovo

+) è veloce

....alternatives...

Maria Luisa Sapino (BDMM 2010)

62
9

2) scelgo il cluster più vicino

+) crea cluster più compatti, ovvero cluster in cui la vicinanza media reciproca tra gli elementi è minima
+) il rappresentante è più rappresentativo

....alternatives...

63
0

Maria Luisa Sapino (BDMM 2010)

3) scelgo il cluster con cardinalità minore

+) porta ad un carico più bilanciato perché distribuisce gli elementi in modo omogeneo: alla fine è più probabile che i cluster così creati siano bilanciati
+) entropia alta, perché distribuendo in maniera uniforme gli elementi c'è equiprobabilità che un elemento finisca in qualsiasi cluster

N.B. se un cluster fosse grandissimo rispetto agli altri, la probabilità che un nuovo elemento finisca lì dentro sarebbe molto più elevata (stessa cosa vale per una eventuale query)

What if we do not have a threshold??

- Find a minimum spanning tree of the input graph
 - $O(N^2)$ edges to $O(N)$ edges

63
1

Maria Luisa Sapino (BDMM 2010)

finora abbiamo preso come ipotesi di avere una threshold per decidere se inserire un elemento in un cluster: come si definisce T ?

se non abbiamo T dobbiamo ragionare sui valori e definire T in base ai dati

1) prendo il grafo iniziale completo pesato in cui per ogni coppia di vertici esiste un arco che li collega e che ha peso pari alla loro somiglianza: ho $O(n)$ nodi e $O(n^2)$ archi
---> di questo albero calcolo MST -> ricopre tutti i vertici ed ha costo minimo

What if we do not have a threshold??

- Find a minimum spanning tree of the input graph
 - $O(N^2)$ edges to $O(N)$ edges
- Remove all edges longer than the average of their neighbors
 - threshold is determined based on the neighborhood

63
2

Maria Luisa Sapino (BDMM 2010)

2) consideriamo per ciascun nodo come si relaziona con i suoi vicini: prendo un arco e guardo il peso dell'arco e dei suoi vicini (ossia degli archi che partono dai nodi collegati dall'arco in questione) -> se l'arco ha peso maggiore della media dei pesi dei suoi vicini, allora è un arco di separazione tra gruppi diversi e quindi lo si elimina -> quindi la threshold non è più unica e globale ma viene calcolata localmente in base ai dati

What if we do not have a threshold??

- Find a minimum spanning tree of the input graph
 - $O(N^2)$ edges to $O(N)$ edges
- Remove all edges longer than the average of its neighbors
 - threshold is determined based on the neighborhood
- Apply connected-components or clique..

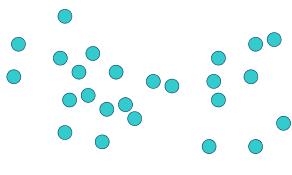
63
3

Maria Luisa Sapino (BDMM 2010)

ora si può applicare il calcolo delle componenti connesse per ottenere i cluster

Max-a-min

- The number of clusters is known, r (say 4)



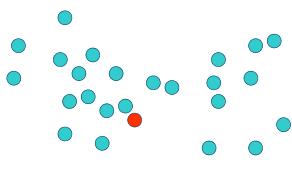
63
4

Maria Luisa Sapino (BDMM 2010)

fino a questo punto abbiamo lavorato sulle distanze guidati dalla threshold; ora invece voglio definire a priori QUANTI devono essere i gruppi

Max-a-min

- Choose a random leader



63
5

Maria Luisa Sapino (BDMM 2010)

1) scelgo a caso un leader

Max-a-min

- Find the furthest point to the leader

Maria Luisa Sapino (BDMM 2010)

63
6

2) il secondo che scelgo è il più lontano possibile dal primo che ho scelto

Max-a-min

- Find the furthest point to the 2 leaders

Maria Luisa Sapino (BDMM 2010)

63
7

3) scelgo il terzo più lontano da entrambi

Max-a-min

- Find the furthest point to the 3 leaders

Maria Luisa Sapino (BDMM 2010)

63
8

4) scelgo il quarto più lontano dai primi 3

Max-a-min

- Assign points to closest leader..

Maria Luisa Sapino (BDMM 2010)

63
9

a questo punto assegno ogni elemento al più vicino leader

K-means (iterative improvement)

- Minimize a “global cost function”

Maria Luisa Sapino (BDMM 2010)

64
0

devo definire una funzione di costo globale:
devo anche tenere conto della compattezza e altre proprietà, non solo della vicinanza con il leader!

ad es. è possibile che un elemento che ho messo in un cluster in realtà fosse più vicino ad un altro cluster

K-means (iterative improvement)

- Minimize a “global cost function”

Maria Luisa Sapino (BDMM 2010)

- ...each item is checked whether moving to another cluster would reduce global cost

64
1

per ogni oggetto del cluster controllo se lo spostamento in un altro cluster ridurrebbe il costo globale

K-means (iterative improvement)

- Minimize a “global cost function”

- ...each item is checked whether moving to another cluster would reduce global cost

64
2

Maria Luisa Sapino (BDMM 2010)

K-means (iterative improvement)

- Minimize a “global cost function”

- ...each item is checked whether moving to another cluster would reduce global cost

64
3

Maria Luisa Sapino (BDMM 2010)

K-means (iterative improvement)

- Minimize a “global cost function”

- ...each item is checked whether moving to another cluster would reduce global cost

64
4

Maria Luisa Sapino (BDMM 2010)

K-means (iterative improvement)

- Minimize a “global cost function”

- ...each item is checked whether moving to another cluster would reduce global cost

64
5

Maria Luisa Sapino (BDMM 2010)

K-means (iterative improvement)

- Minimize a “global cost function”

- ...each item is checked whether moving to another cluster would reduce global cost

64
6

Maria Luisa Sapino (BDMM 2010)

K-means (iterative improvement)

- Minimize a “global cost function”

- ...each item is checked whether moving to another cluster would reduce global cost

64
7

Maria Luisa Sapino (BDMM 2010)

K-means (iterative improvement)

- Minimize a “global cost function”

K-means (iterative improvement)

- Minimize a “global cost function”

What are the possible criteria??

- Compactness (minimize root-mean-square)

RMSE_i lo calcolo per ogni cluster, quindi mi dà solo un valore locale: voglio una misura globale che tenga conto della qualità (compattezza) di TUTTI i cluster

RMSE è la media dei valori degli RMSE_i ed è il valore che andiamo a minimizzare

è possibile minimizzare la funzione root-mean-square che misura la compattezza:
per ogni punto del dominio

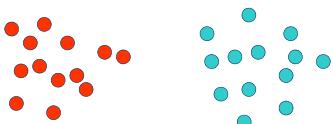
$F^*(z_i)$ è il leader del cluster

quindi la prima formula rappresenta l'errore che faccio nel fare i confronti se considero il leader del cluster anziché l'elemento iniziale -> la sommatoria di questi errori quadrati diviso per N è l'errore quadratico medio a livello di ciascun cluster (m è il # dei cluster)

più è basso questo valore più il gruppo è compatto e il leader rappresenta bene tutti gli altri

What are the possible criteria???

- Evenly sized clusters (maximize entropy)



$$H(X) \equiv \sum_{x \in \mathcal{A}_X} P(x) \log \frac{1}{P(x)}$$

Maria Luisa Sapino (BDMM 2010)

65
1

altro possibile criterio per cluster ottimali:

cluster di dimensioni omogenee -> questo significa massimizzare l'ENTROPIA, ovvero l'appartenenza di un elemento a ciascun cluster è il più possibile equiprobabile

What if we do not have distances???

- ...we need to learn from
 - user feedback or
 - user access patterns!!

Maria Luisa Sapino (BDMM 2010)

65
2

se non conosciamo le distanze?
allora dobbiamo prendere informazioni da

- * feedback dell'utente (blackbox)
- * pattern di accesso dell'utente

Confidence clustering...

- Assumption:
 - each cluster can have maximum r objects
- ...keeps adapting the clusters to user access pattern

Maria Luisa Sapino (BDMM 2010)

65
3

CONFIDENCE CLUSTERING:

1) assumiamo di avere una cardinalità massima per ogni cluster, ad es. ogni cluster può contenere al massimo r oggetti

2) adattiamo i clusters in base al pattern di accesso dell'utente

Confidence clustering...

- Assign each object o_i to cluster C_j with random confidence ($0 \leq \text{conf}(i,j) \leq 10$)

65
4

Maria Luisa Sapino (BDMM 2010)

Confidence clustering...

- Assign each object o_i to cluster C_j with random confidence ($0 \leq \text{conf}(i,j) \leq 10$)
- If user accesses o_a and o_b , then

65
5

Maria Luisa Sapino (BDMM 2010)

Confidence clustering...

- Assign each object o_i to cluster C_j with random confidence ($0 \leq \text{conf}(i,j) \leq 10$)
- If user accesses o_a and o_b , then
 - if o_a and o_b are in the same cluster C_j then
 - $\text{conf}(a,j)++$
 - $\text{conf}(b,j)++$

65
6

Maria Luisa Sapino (BDMM 2010)

- inizialmente si assegna ogni oggetto o_i al cluster C_j con una certa confidenza $\text{conf}(i, j)$ scelta a caso tra 0 e 10

- se l'utente accede sia ad O_a che ad O_b e questi sono nello stesso cluster, allora verosimilmente possiamo aumentare la confidenza che i due oggetti vadano tenuti nello stesso cluster, quindi aumentiamo la confidenza di entrambi rispetto al cluster C_j :

* $\text{conf}(a, j)++$
 * $\text{conf}(b, j)++$

Confidence clustering...

- Assign each object o_i to cluster C_j with random confidence ($0 \leq \text{conf}(i,j) \leq 10$)
- If user accesses o_a and o_b , then
 - if o_a is in cluster C_i and o_b is in cluster C_j then

65
7

Maria Luisa Sapino (BDMM 2010)

Confidence clustering...

- Assign each object o_i to cluster C_j with random confidence ($0 \leq \text{conf}(i,j) \leq 10$)
- If user accesses o_a and o_b , then
 - if o_a is in cluster C_i and o_b is in cluster C_j then
 - if $\text{conf}(a,i) > 1$ and $\text{conf}(b,j) > 1$ then
 - $\text{conf}(a,i) --$
 - $\text{conf}(b,j) --$

65
8

Maria Luisa Sapino (BDMM 2010)

- se l'utente accede sia ad O_a che ad O_b e questi sono in due cluster diversi, con O_a che appartiene a C_i e O_b che appartiene a C_j , allora agiamo in modo diverso a seconda del valore di confidenza attuale dei due oggetti rispetto al loro cluster di appartenenza:

1) se hanno entrambi $\text{conf} > 1$, verosimilmente possiamo diminuire la confidenza che i due oggetti vadano tenuti in due cluster diversi, quindi dicrementiamo la confidenza di entrambi rispetto al proprio cluster di appartenenza:

* $\text{conf}(a, i) --$
* $\text{conf}(b, j) --$

2) se entrambi hanno $\text{conf} = 1$, verosimilmente i due oggetti dovrebbero appartenere allo stesso cluster, quindi proviamo a vedere come evolve la situazione ponendo O_b nel cluster C_i (con confidenza minima)

* $\text{conf}(b, i) = 1$

e spostando un oggetto O_c da C_i a C_j (per lasciare il posto a O_b), ad es. quello che ha confidenza minore

* $\text{conf}(c, j) = 1$

Confidence clustering...

- Assign each object o_i to cluster C_j with random confidence ($0 \leq \text{conf}(i,j) \leq 10$)
- If user accesses o_a and o_b , then
 - if o_a is in cluster C_i and o_b is in cluster C_j then
 - if $\text{conf}(a,i) == 1$ and $\text{conf}(b,j) == 1$ then
 - $\text{conf}(b,i) = 1$ (move o_b from C_j to C_i)
 - $\text{conf}(c,j) = 1$ (move some o_c from C_i to C_j)

65
9

Maria Luisa Sapino (BDMM 2010)

Confidence clustering...

- Assign each object o_i to cluster C_j with random confidence ($0 \leq \text{conf}(i,j) \leq 10$)
- If user accesses o_a and o_b , then
 - if o_a is in cluster C_i and o_b is in cluster C_j then
 - if $\text{conf}(a,i) > 1$ and $\text{conf}(b,j) == 1$ then
 - there exists $\text{conf}(c,i) == 1$
 - $\text{conf}(b,i) = 1$ (move o_b from C_j to C_i)
 - $\text{conf}(c,j) = 1$ (move o_c from C_i to C_j)

Maria Luisa Sapino (BDMM 2010)

66
0

2) se uno dei due ha $\text{conf}=1$, ad es. O_b , provo ad effettuare lo scambio con un elemento di C_i che abbia $\text{conf} = 1$

- se esiste un tale elemento O_c , proviamo a vedere come evolve la situazione ponendo O_b nel cluster C_i (con confidenza minima)

$$* \text{conf}(b, i) = 1$$

e spostando l'oggetto O_c da C_i a C_j (per lasciare il posto a O_b)

$$* \text{conf}(c, j) = 1$$

Confidence clustering...

- Assign each object o_i to cluster C_j with random confidence ($0 \leq \text{conf}(i,j) \leq 10$)
- If user accesses o_a and o_b , then
 - if o_a is in cluster C_i and o_b is in cluster C_j then
 - if $\text{conf}(a,i) > 1$ and $\text{conf}(b,j) == 1$ then
 - there **does not** exist $\text{conf}(c,i) == 1$
 - $\text{conf}(a,i) = 1$

Maria Luisa Sapino (BDMM 2010)

66
1

- se non esiste un tale elemento O_c , allora diminuiamo semplicemente la confidenza che O_a debba stare in C_i

$$* \text{conf}(a, i) = 1$$

Adaptive clustering...

- What if we do not know the number of clusters????
- ...keeps adapting the clusters to user access pattern

Maria Luisa Sapino (BDMM 2010)

66
2

se non conosciamo il numero di cluster da formare, possiamo comunque usare algoritmi che effettuano raffinamenti successivi per ADATTARE i cluster al pattern di accesso dell'utente

Adaptive clustering...



- start with a random assignment of objects to a line

Maria Luisa Sapino (BDMM 2010)

Adaptive clustering...



- If a set of objects are accessed together...

Maria Luisa Sapino (BDMM 2010)

Adaptive clustering...



- ..pull the objects closer to their average point..

Maria Luisa Sapino (BDMM 2010)

1) inizialmente si assegnano randomicamente gli oggetti disponendoli su una linea

2) ogni volta che un insieme di oggetti viene acceduto contemporaneamente, tutti gli oggetti dell'insieme vengono avvicinati rispetto al loro punto medio sulla linea

Adaptive clustering...



- ..choose another random set of objects..

Maria Luisa Sapino (BDMM 2010)

66
6

Adaptive clustering...



- ..push these away from their average point..

Maria Luisa Sapino (BDMM 2010)

66
7

Adaptive clustering...



- ...over time...
- similar objects will come closer...
- different objects will get apart...

Maria Luisa Sapino (BDMM 2010)

66
8

3) per fare in modo che gli oggetti non rimangano troppo addensati al centro della linea, si sceglie un insieme random di elementi che vengono invece ALLONTANATI dal loro punto medio

in questo modo, dopo un certo numero di iterazioni si avrà che:

- * gli oggetti simili sono raggruppati insieme sulla linea
- * gli oggetti diversi sono separati tra loro sulla linea

Support vector machines

- Given a feature space of objects and example labels, find surfaces that separate the data into clusters

66
9

JMM 2010
Figures from Burges 98

altro approccio per raggruppare oggetti: SVM

dato uno spazio in cui sono rappresentati degli oggetti, e in cui a ciascun oggetto è assegnata una certa ETICHETTA, vogliamo trovare delle superfici che separino i dati in cluster tali che elementi nello stesso cluster abbiano la stessa etichetta

Support vector machines

- Given a feature space of objects and example labels, find surfaces that separate the data into clusters

67
0

Maria Luisa Sapino (IISMM 2010)
Figures from Burges 98

ogni oggetto dello spazio può essere visto come una coppia $\langle \vec{x}, y \rangle$, dove $\vec{x} \in \mathbb{R}^n$ è un vettore nello spazio n-dimensionale e y è l'etichetta associata all'oggetto

l'obiettivo della SVM è quello di partizionare lo spazio in regioni tali che ogni regione contenga oggetti con la stessa etichetta

Support vector machines

- Given a feature space of objects and example labels, find surfaces that separate the data into clusters

67
1

Maria Luisa Sapino (IISMM 2010)
Figures from Burges 98

se non c'è una linea che separa gli oggetti, invece che trovare una curva che si adatti ai dati, si usa una FUNZIONE KERNEL per mappare i dati in uno spazio diverso in cui si possa usare un (iper)piano per separarli

How can we estimate the number of clusters in a database?

- ...some clustering algorithms require the expected number of clusters as input

67
2

Maria Luisa Sapino (BDMM 2010)

abbiamo bisogno di metodi per stimare il numero di cluster perché alcuni algoritmi richiedono questo parametro come input per poter funzionare!

Si vuole stimare un buon numero di cluster che il DB deve avere.

Buon numero di cluster: numero di cluster che mi permette di avere buone proprietà (compattezza, entropia, ecc.)

How can we estimate the number of clusters in a database?

- ...some clustering algorithms require the expected number of clusters as input
 - covering

covering(o_i, o_j) = $\sum_{k=1..n} p(k | o_i)p(o_j | k)$

importance of feature f_k in o_i

probability that o_j is a document having feature f_k

Maria Luisa Sapino (BDMM 2010)

67
3

—
—

$P(k | O_i)$ = l'importanza che la feature k ha per l'oggetto O_i

il COVERING è una grandezza definita su coppie di oggetti che correla le features presenti in un oggetto a quelle che sono presenti nell'altro: esprime un concetto di somiglianza, quanto l'oggetto i ricopre l'oggetto j

per fare questo considero una ad una tutte le features dell'oggetto i, e guardo in qualche misura le features che sono presenti in i siano caratterizzanti per l'oggetto j

per tutte le features dei miei oggetti, considero il prodotto delle due probabilità condizionate: la prima è la P che dato l'oggetto i la feature k sia importante per quell'oggetto (ovvero abbia un valore alto), la seconda è la P che dato l'oggetto j la stessa feature k sia altrettanto caratterizzante

How can we estimate the number of clusters in a database?

- ...some clustering algorithms require the expected number of clusters as input
 - covering

$$\text{covering}(o_i, o_j) = \sum_{k=1..n} p(k | o_i)p(o_j | k)$$


probability that o_j is a document having feature f_k

Maria Luisa Sapino (BDMM 2010)

67
4

—
—

How can we estimate the number of clusters in a database?

- ...some clustering algorithms require the expected number of clusters as input
- covering

$$\text{covering}(o_i, o_j) = \sum_{k=1..n} p(k | o_i) p(o_j | k)$$

67
5



Maria Luisa Sapino (BDMM 2010)

sapendo che la feature k è importante per O_i, se io faccio una ricerca su tutti gli oggetti che soddisfano la feature k, qual'è la P che mi venga restituito O_j?

gli oggetti meglio ricoperti sono quelli restituiti dalla ricerca

How can we estimate the number of clusters in a database?

$$\text{covering}(o_i, o_j) = \sum_{k=1..n} p(k | o_i) p(o_j | k)$$

importance of feature f_k in o_i probability that o_j is a document having feature f_k

- Suppose the database is a perfect cluster
 - features are uniformly distributed
 - all documents are equally likely to be selected

67
6



Maria Luisa Sapino (BDMM 2010)

$\text{covering}(O_i, O_j) = \text{quanto le features di } O_i \text{ sono ben caratterizzanti per } O_j$

inizialmente supponiamo una situazione ideale in cui il DB è un unico cluster perfetto, ovvero

- tutte le features sono distribuite in modo uniforme
- gli elementi sono tutti simili tra loro

How can we estimate the number of clusters in a database?

$$\text{covering}(o_i, o_j) = \sum_{k=1..n} p(k | o_i) p(o_j | k)$$

importance of feature f_k in o_i probability that o_j is a document having feature f_k

- Suppose the database is a single cluster

$$\text{covering}(o_i, o_j) = \sum_{k=1..n} \frac{1}{n} \frac{1}{D} = n \frac{1}{n} \frac{1}{D} = \frac{1}{D}$$

67
7

Maria Luisa Sapino (BDMM 2010)

n = numero delle features

D = numero di oggetti

quindi per qualsiasi k la probabilità di selezionare l'oggetto è la stessa e tutti i documenti hanno la stessa P di essere selezionati:

allora $P(k | O_i) = 1/n$ perché tutte le features hanno la stessa importanza, e

$P(k | O_j) = 1/D$ perché ogni documento ha la stessa probabilità di essere selezionato

allora il covering di ogni coppia di oggetti è $1/D$

How can we estimate the number of clusters in a database?

- Suppose the database is a single cluster

$$\text{covering}(o_i, o_j) = \sum_{k=1..n} \frac{1}{n} \frac{1}{D} = n \frac{1}{n} \frac{1}{D} = \frac{1}{D}$$

- Let's sum up all self-coverings, then

$$\sum_{o_i} \text{covering}(o_i, o_i) = \sum_D \frac{1}{D} = 1$$

67
8

Maria Luisa Sapino (BDMM 2010)

se considero complessivamente il covering di ciascun oggetto con se stesso, avrò D oggetti il cui self-covering = $1/D$
N.B. $1/D$ è il covering tra due oggetti e quindi anche tra un oggetto e sé stesso

quindi complessivamente la somma dei self-covering è = 1

How can we estimate the number of clusters in a database?

- Suppose the database is a single cluster

$$\text{covering}(o_i, o_j) = \sum_{k=1..n} \frac{1}{n} \frac{1}{D} = n \frac{1}{n} \frac{1}{D} = \frac{1}{D}$$

- Let's sum up all self-coverings, then

$$\sum_{o_i} \text{covering}(o_i, o_i) = \sum_D \frac{1}{D} = 1$$

68
0

Maria Luisa Sapino (BDMM 2010)

ora, abbiamo due clusters C1 con m1 oggetti e C2 con m2 oggetti: quale è il covering globale del sistema? -> è il covering totale tra la somma di due sommatorie, quella di C1 e quella di C2

$$\text{covering}(i, j) = \sum_{O_i \in C1} \sum_{O_j \in C2} (1/n * 1/m1) + \sum_{O_j \in C2} \sum_{O_i \in C1} (1/n * 1/m2) = \sum_{O_i \in C1} (n * 1/n * 1/m1) + \sum_{O_j \in C2} (n * 1/n * 1/m2) = m1 * 1/m1 + m2 * 1/m2 = 1 + 1 = 2$$

How can we estimate the number of clusters in a database?

$$\sum_{o_i} \text{covering}(o_i, o_i) = p$$

There are approximately p clusters

68
1

Maria Luisa Sapino (BDMM 2010)

il covering globale per ciascun oggetto con sé stesso all'interno di un unico cluster ideale è 1

* 1 singolo cluster
* 1 è il valore del covering globale

quindi partendo dall'hp di avere p cluster ideali con distribuzione uniforme, il covering totale degli oggetti rispetto a se stessi dà esattamente il valore di p
--> se non si conosce a priori il # di cluster potenzialmente riconoscibili in un DB, una stima di questo numero può essere fatta calcolando il covering totale degli oggetti

N.B. il calcolo del covering totale è più complicato perché va fatto utilizzando la prima formula (noi l'abbiamo istanziata al caso di cluster uniformi)

posso stimare il numero di cluster con questo valore

Use of clusters (prune search space)

- ...eliminate clusters based on their representatives

Maria Luisa Sapino (BDMM 2010)

Use of clusters Binary independent features

- Each document is a binary vector
- Documents are organized into clusters
- Each cluster has a representative

Maria Luisa Sapino (BDMM 2010)

Use of clusters Binary independent features

- Each document is a binary vector
- Documents are organized into clusters
- Each cluster has a representative
- Goal: for each cluster, estimate # of documents having t or more matching keywords with a query with k keywords

Maria Luisa Sapino (BDMM 2010)

a cosa ci servono i cluster?

* possiamo utilizzarli per fare pruning ed eliminare dallo spazio di ricerca i cluster il cui rappresentante non soddisfa i requisiti che stiamo cercando
 --> è un pruning diverso da quello degli indici: infatti con gli indici lavoriamo in uno spazio metrico e quindi facendo il pruning siamo sicuri di eliminare dati che non ci interessano -> con il clustering invece si possono perdere dei dati perché ci si basa sul leader per confrontare e prendere decisioni, e si possono restituire false hits

altro uso dei clusters:

- consideriamo oggetti che sono vettori di features rappresentate come valori binari (solo 0 o 1 risp. se la feature è presente o no nell'oggetto)
 --> è sempre possibile passare da una rappresentazione complessa a una rappresentazione binaria: 0 rimane 0, qualsiasi peso > 0 diventa 1

- documenti organizzati in cluster che comprendono oggetti simili tra loro

- ciascun cluster è associato ad un rappresentante

per ciascun cluster vogliamo stimare il # di documenti che abbiano almeno t chiavi che matchano con quelli di una keyword che ne contiene k

se ho documenti definiti su vettori di 3000 elementi, ho keyword di 30 -> quanti documenti nel cluster soddisfano almeno 20 delle 30 keyword date?

Use of clusters Binary independent features

- Each document is a binary vector
- Documents are organized into clusters
- Each cluster has a representative

$$o_i = \langle f_{i,1}, f_{i,2}, \dots, f_{i,n} \rangle; \quad R_O = \langle r_1, r_2, \dots, r_n \rangle = \frac{o_i}{|O|}$$

Probability that a document in the cluster has this keyword

$$q = \langle 1, 1, 1, \dots, 1, 0, 0, \dots, 0 \rangle; \text{with } k \text{ 1s}$$

Maria Luisa Sapino (BDMM 2010)

68
5

- ogni oggetto i è rappresentato da un vettore lungo n

- R_O è la rappresentazione che riassume in forma compatta tutti gli oggetti del cluster:

r_i = frequenza con cui la feature i è presente nel dominio

- la query è rappresentata in formato binario e contiene k 1s

Use of clusters Binary independent features

- Each document is a binary vector
- Documents are organized into clusters
- Each cluster has a representative

$$o_i = \langle f_{i,1}, f_{i,2}, \dots, f_{i,n} \rangle; \quad o^k = \langle f_1, f_2, \dots, f_k \rangle; \\ R_O = \langle r_1, r_2, \dots, r_n \rangle = \frac{o_i}{|O|}$$

$$q = \langle 1, 1, 1, \dots, 1, 0, 0, \dots, 0 \rangle; \text{with } k \text{ 1s}$$

Maria Luisa Sapino (BDMM 2010)

68
6

considero un generico oggetto che contiene esattamente le k keyword che cerco: qual'è la probabilità che quell'oggetto appartenga al cluster?

devo considerare uno ad uno gli elementi e vedere come si rapportino alla descrizione compatta del cluster!

per calcolare la P che l'oggetto appartenga al cluster
considero ad una ad una le features e moltiplico la probabilità di trovare quella feature nella query (se la volevo) e di non trovarla (se non la volevo)

--> 1. la P di trovare la feature r_j se la volevo
--> 2. la P di non trovare la feature se non la volevo

Use of clusters Binary independent features

$$o_i = \langle f_{i,1}, f_{i,2}, \dots, f_{i,n} \rangle; \quad o^k = \langle f_1, f_2, \dots, f_k \rangle; \\ R_O = \langle r_1, r_2, \dots, r_n \rangle = \frac{o_i}{|O|}$$

$$p(o^k \in O) = \prod_{j=1}^k (r_j)^{f_j} (1 - r_j)^{1-f_j}$$

Maria Luisa Sapino (BDMM 2010)

68
7

mi consente di stimare la P che un certo oggetto con certe componenti stia nel DB

Use of clusters Binary independent features

$$o_i = \langle f_{i,1}, f_{i,2}, \dots, f_{i,n} \rangle; \quad o^k = \langle f_1, f_2, \dots, f_k \rangle;$$

$$R_O = \langle r_1, r_2, \dots, r_n \rangle = \sum_{o_i \in O} o_i$$

$$num(t, Q) = \sum_{o^k \text{ with } t \text{ 1s}} (p(o^k \in Q))$$

68
8

Maria Luisa Sapino (BDMM 2010)

noi vogliamo stimare, data una query che contiene k 1, quanti oggetti all'interno del cluster ne soddisfano almeno t:

per ogni oggetto che ha almeno k elementi sappiamo calcolare la P che sia presente nel DB: se facciamo la sommatoria estesa a tutti i vettori lunghi k che contengono almeno t volte 1 (oggetti che, se presenti, soddisfano almeno t keyword della query)

--> stimo di poter trovare almeno num(t, Q) che contengono almeno t elementi di Q

Use of clusters Non-binary, independent features

- Each document is a non-binary vector
- Documents are organized into clusters
- Each cluster has a representative
- Goal: for each cluster, find the probability that one object in the cluster will be more than S similar to the query

68
9

Maria Luisa Sapino (BDMM 2010)

quale probabilità che ci sia almeno un oggetto nel cluster che sia simile almeno S alla query

Use of clusters Non-binary independent features

$$o_i = \langle f_{i,1}, f_{i,2}, \dots, f_{i,n} \rangle; \quad o^k = \langle f_1, f_2, \dots, f_k \rangle;$$

69
0

Maria Luisa Sapino (BDMM 2010)

ora abbiamo features con pesi reali (non più binari)

Use of clusters Non-binary independent features

$$o_i = \langle f_{i,1}, f_{i,2}, \dots, f_{i,n} \rangle; \quad o^k = \langle f_1, f_2, \dots, f_k \rangle;$$

$$R_O = \langle [r_1, w_1], [r_2, w_2], \dots, [r_n, w_n] \rangle$$

$$q = \langle q_1, q_2, \dots, q_k \rangle$$

Probability that a document in the cluster has this keyword

69
1

Maria Luisa Sapino (BDMM 2010)

rappresentante del cluster ora deve darmi dell'informazione in più:

oltre all'informazione data da r_i , ovvero la probabilità che un certo documento del cluster contenga la corrispondente keyword i , devo conservare il peso medio associato alla corrispondente keyword i

Use of clusters Non-binary independent features

$$o_i = \langle f_{i,1}, f_{i,2}, \dots, f_{i,n} \rangle; \quad o^k = \langle f_1, f_2, \dots, f_k \rangle;$$

$$R_O = \langle [r_1, w_1], [r_2, w_2], \dots, [r_n, w_n] \rangle$$

$$q = \langle q_1, q_2, \dots, q_k \rangle$$

The average weight of the keyword in the documents that have this keyword

69
2

Maria Luisa Sapino (BDMM 2010)

ho query q che contiene k keyword

vogliamo calcolare la somiglianza usando il dot product:

definiamo $\text{cont}(j, Q)$ come il contributo che la keyword i -esima dà al soddisfacimento della query

--> prodotto del peso medio della componente i -esima per la corrispondente componente della query, con la probabilità associata a quella componente

--> per ciascun componente i ho $P = r_i$ di contribuire alla misura di somiglianza con $w_i * q_i$

Use of clusters Non-binary independent features

$$o_i = \langle f_{i,1}, f_{i,2}, \dots, f_{i,n} \rangle; \quad o^k = \langle f_1, f_2, \dots, f_k \rangle;$$

$$R_O = \langle [r_1, w_1], [r_2, w_2], \dots, [r_n, w_n] \rangle$$

$$q = \langle q_1, q_2, \dots, q_k \rangle$$

$\text{cont}(i, Q) = w_i q_i$; with r_i probability

69
3

Maria Luisa Sapino (BDMM 2010)

Use of clusters Non-binary independent features

$$o_i = \langle f_{i,1}, f_{i,2}, \dots, f_{i,n} \rangle; \quad o^k = \langle f_1, f_2, \dots, f_k \rangle;$$

$$R_O = \langle [r_1, w_1], [r_2, w_2], \dots, [r_n, w_n] \rangle$$

$$q = \langle q_1, q_2, \dots, q_k \rangle$$

$$p(sim(O, Q) = s) = \text{coef} \left(x^s, \prod_{i=1}^k (r_i x^{w_i q_i} + (1 - r_i)) \right)$$

69
4

Maria Luisa Sapino (BDMM 2010)

supponiamo di avere una query Q con 2 keywords:
vogliamo verificare quale è la P che nel cluster si trovi
una somiglianza che raggiunga almeno un certo valore
S rispetto alla presenza delle 2 keyword >> devo
considerare 2 aspetti:

- 1) la P di trovare all'interno dei documenti del cluster
le 2 keyword
- 2) le 2 keyword trovate devono avere un peso tale
che la somma dei loro contributi sia almeno S

N.B. possiamo usare la produttoria perché le features
sono indipendenti

Use of clusters Non-binary independent features

$$o_i = \langle f_{i,1}, f_{i,2}, \dots, f_{i,n} \rangle; \quad o^k = \langle f_1, f_2, \dots, f_k \rangle;$$

$$R_O = \langle [r_1, w_1], [r_2, w_2], \dots, [r_n, w_n] \rangle$$

$$q = \langle q_1, q_2, \dots, q_k \rangle$$

$$p(sim(O, Q) = s) = \text{coef} \left(x^s, \prod_{i=1}^k (r_i x^{w_i q_i} + (1 - r_i)) \right)$$

69
5

Maria Luisa Sapino (BDMM 2010)

se ci sono le features che cerco, il contributo che
queste danno deve essere sommato:

la rappresentazione polinomiale mi permette di
gestire facilmente prodotto e somma

$(1 - r_i)$ = P che nell'oggetto non ci sia la keyword i

---> la P che cerchiamo è data dal coefficiente del
termine con grado S!

What if features are not independent?

- Metric spaces assume that features are independent (orthogonal to each other)
- ...what if they are not?

69
8

Maria Luisa Sapino (BDMM 2010)

non è detto che i termini siano sempre tra loro
indipendenti: può esserci correlazione tra gli elementi

anche con gli indici assumevamo che le features
fossero ortogonali tra loro

Latent Semantic Indexing

- Used for hidden (latent) concepts in a given collection
 - mostly for text collections (cosine similarity!)
- Let us have
 - $|O|$ objects
 - Each object o is represented with a vector of size $|V|$ (number of features)

69
9

Maria Luisa Sapino (BDMM 2010)

===== Latent Semantic Indexing =====

tecniche che hanno l'obiettivo di far emergere dai documenti le relazioni semantiche latenti tra i concetti
 --> es. "papà e mamma" ha un significato diverso da "papà" e "mamma" quando occorrono separati

se i termini non sono tra loro indipendenti, si trasforma lo spazio in cui le features sono non indipendenti in un nuovo spazio in cui le dimensioni sono tra loro indipendenti: i dati possono ora essere proiettati in questo nuovo spazio

abbiamo $|O|$ oggetti ciascun dei quali è rappresentato da un vettore lungo $|V|$

Document-feature vector

OF =	1	2	V	feature value
	1				
	2				
				
	O				

70
0

Maria Luisa Sapino (BDMM 2010)

rappresentazione compatta del DB:

ogni riga è un vettore documento che ha $|V|$ documenti
 e per ogni documento mi dice quali sono i valori per ogni feature

nell'altro senso posso leggerla come:
 in quali documenti e con quali pesi è presente la feature k

How can we use this matrix?

- This matrix is the database!!!
- Can we use it to find
 - object-object similarities?
 - feature-feature correlation?
 - independent concepts in the collection?
- Can we use it for efficient indexing?

70
1

Maria Luisa Sapino (BDMM 2010)

posso usare questa matrice per estrarre info su:

- 1) somiglianza tra oggetti
- 2) correlazione tra features
- 3) possibilità di estrarre concetti indipendenti tra loro da questa collezione di documenti, ovvero concetti che non possono essere definiti l'uno nei termini dell'altro (es. altezza e colore dei capelli, mentre ad es. altezza e peso sono correlati)

Obj-feature X feature-obj

$$\text{OF} \times \text{OF}^T = \begin{matrix} & 1 & 2 & \dots & |\mathcal{M}| \\ \begin{matrix} 1 \\ 2 \\ \vdots \\ |\mathcal{O}| \end{matrix} & \begin{matrix} | \\ | \\ \vdots \\ | \end{matrix} & \times & \begin{matrix} | \\ | \\ \vdots \\ | \end{matrix} & \begin{matrix} 1 & 2 & \dots & |\mathcal{O}| \\ 1 & 2 & \vdots & |\mathcal{M}| \end{matrix} \end{matrix}$$

70
2

Maria Luisa Sapino (BDMM 2010)

== CALCOLARE SOMIGLIANZA TRA OGGETTI ==

moltiplico la matrice O-F per la matrice trasposta,
ovvero la matrice F-O in cui scambio le righe con le
colonne

Obj-feature X feature-obj

$$\text{OF} \times \text{OF}^T = \begin{matrix} & 1 & 2 & \dots & |\mathcal{M}| \\ \begin{matrix} 1 \\ 2 \\ \vdots \\ |\mathcal{O}| \end{matrix} & \begin{matrix} | \\ | \\ \vdots \\ | \end{matrix} & \times & \begin{matrix} | \\ | \\ \vdots \\ | \end{matrix} & \begin{matrix} 1 & 2 & \dots & |\mathcal{O}| \\ 1 & 2 & \vdots & |\mathcal{M}| \end{matrix} \end{matrix}$$

70
3

vector multiplication (dot)

Maria Luisa Sapino (BDMM 2010)

calcolo prodotto di queste due matrici: alla fine avrò
matrice $|\mathcal{O}| \times |\mathcal{O}|$

Obj-feature X feature-obj

$$\text{OF} \times \text{OF}^T = \begin{matrix} & 1 & 2 & \dots & |\mathcal{M}| \\ \begin{matrix} 1 \\ 2 \\ \vdots \\ |\mathcal{O}| \end{matrix} & \begin{matrix} | \\ | \\ \vdots \\ | \end{matrix} & \times & \begin{matrix} | \\ | \\ \vdots \\ | \end{matrix} & \begin{matrix} 1 & 2 & \dots & |\mathcal{O}| \\ 1 & 2 & \vdots & |\mathcal{M}| \end{matrix} \end{matrix}$$

70
4

vector multiplication (dot)

Maria Luisa Sapino (BDMM 2010)

Obj-feature X feature-obj

in ogni cella (i, j) della matrice risultante ho la somiglianza tra l'oggetto i e l'oggetto j

Obj-obj similarity matrix!!!!

OF x OF! = OO =

70
6

== CALCOLARE CORRELAZIONE TRA FEATURES ==

per trovare correlazione tra features faccio la stessa cosa ma moltiplico la trasposta per la matrice:
(i, j) mi dice quanto la feature i è correlata alla feature j
-> mi dice quanto i e j concorrono a definire gli stessi documenti

Feature-feature correl. matrix!!!!

FO x FO = FF =

1 2

M

V V

1, 2 3, 4

Media | Luisa Sapino (BDMM 2010)

Singular valued decomposition

OF =

Maria Luisa Sapino (BDMM 2010)

70
8

Singular valued decomposition

OF =

OC X CC CV

Maria Luisa Sapino (BDMM 2010)

70
9

Singular valued decomposition

OF =

OC X CC CV

71 columns linearly independent (column orthonormal)
0

Maria Luisa Sapino (BDMM 2010)

===== ESTRARRE CONCETTI INDIPENDENTI =====

data la matrice OF, applico una trasformazione di algebra lineare (calcolo matriciale)

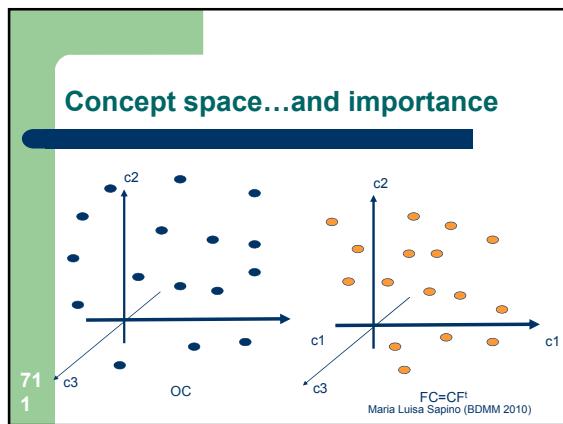
si decompone la matrice di partenza in 3 matrici:
la matrice centrale CC è matrice diagonale che

- * ha valori diversi da 0 solo sulla diagonale
- * i valori sulla diagonale sono disposti in ordine decrescente dalla posizione (1, 1)

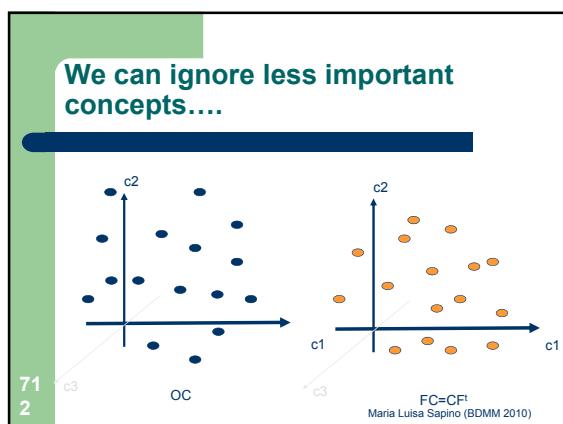
i ICI elementi di questa matrice sono i ICI concetti indipendenti, con $|CI| \leq |v|$

OC = mi dà l'interpretazione degli oggetti nel nuovo spazio

CV⁻¹ = come si definiscono le features nei termini dei nuovi concetti (mi serve per calcolare le distanze)

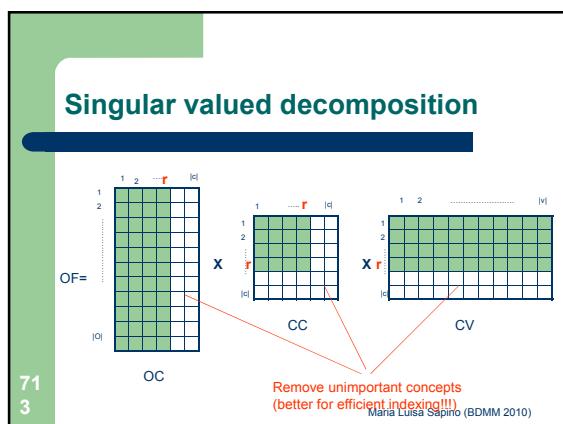


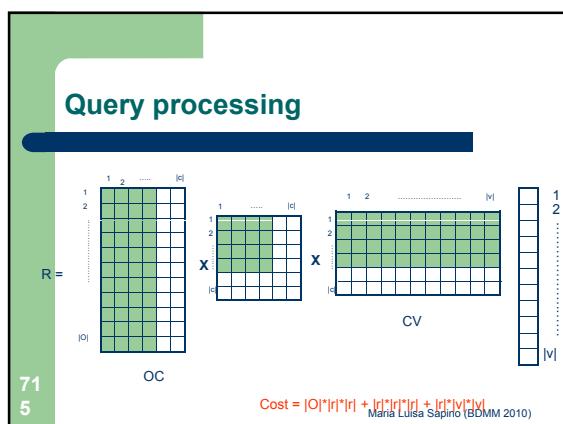
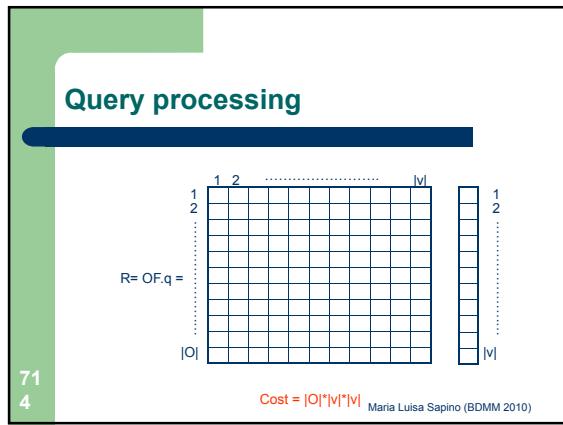
sulla matrice diagonale, le dimensioni ultime danno contributo inferiore alla definizione dell'oggetto, portano meno dettaglio: se voglio ridurre le dimensioni, posso ignorare gli ultimi concetti perché sono quelli meno informativi -> se trascurati hanno impatto meno grave sulla valutazione della somiglianza!



IOI x lvl x lvl è il costo del confronto nello spazio originale

se voglio stimare la distanza nello spazio ridotto e tengo solo R componenti su ICI < lvl, il costo si riduce!





```
C = {01, 02, 03}
* 01 = <3, 0, 2, 5, 7>
* 02 = <0, 4, 1, 3, 5>
* 03 = <2, 0, 0, 0, 4>
```

```
R_C = {[2/3, 5/3], [1/3, 4/3], [2/3, 3/3], [2/3, 8/3], [3/3, 16/3]}
```

```
Q = <0, 1, 0, 0, 1>
```

```
S = 1
```

```
calcolo il polinomio:
```

```
per k = 2 -> P2 = (1/3)*x^(4/3*1) + (1 - 1/3) = (1/3)*x^(4/3) + 2/3
per k = 5 -> P5 = (3/3)*x^(16/3*1) + (1 - 3/3) = x^(16/3)
```

a questo punto il prodotto P2 * P5 mi dà il polinomio cercato:

P2 * P5 = [(1/3)*x^(4/3) + 2/3] * x^(16/3) = (1/3)*x^(20/3) + (2/3)*x^(16/3)

il primo termine rappresenta il caso in cui siano presenti entrambe le keyword, il secondo invece cattura il caso in cui la prima keyword è assente

con P = 1/3 avrò somiglianza 20/3 (quando sono presenti entrambe le features)
con P = 2/3 avrò somiglianza 16/3 (quando è presente solo la feature 2)

quando facciamo una query e otteniamo dei risultati, come valutiamo se i risultati sono conformi alle aspettative dell'utente?
---> il CLUSTERING e l'hashing introducono degli errori!

Use of clusters (prune search space)

71
6

- ...eliminate clusters based on their representatives

Maria Luisa Sapino (BDMM 2010)

abbiamo visto che eliminare i cluster in base ai rappresentanti può introdurre errori

Evaluation of clustering methods

71
7

Maria Luisa Sapino (BDMM 2010)

gli oggetti del DB sono clusterizzati in modo tale che

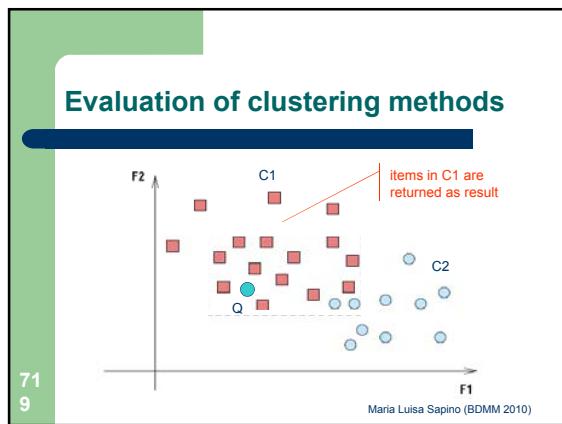
- * i quadratini rossi sono parte del cluster C1
- * i pallini azzurri sono parte del cluster C2

Evaluation of clustering methods

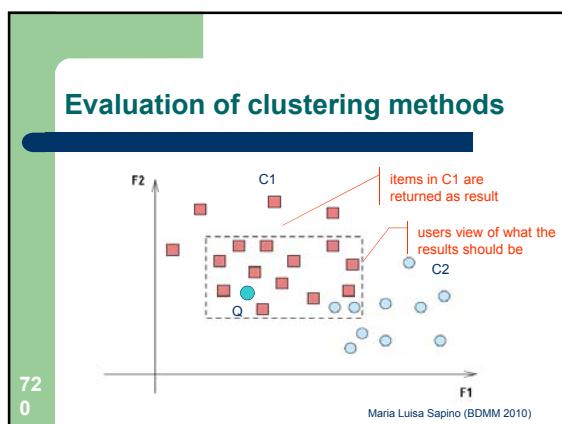
71
8

Maria Luisa Sapino (BDMM 2010)

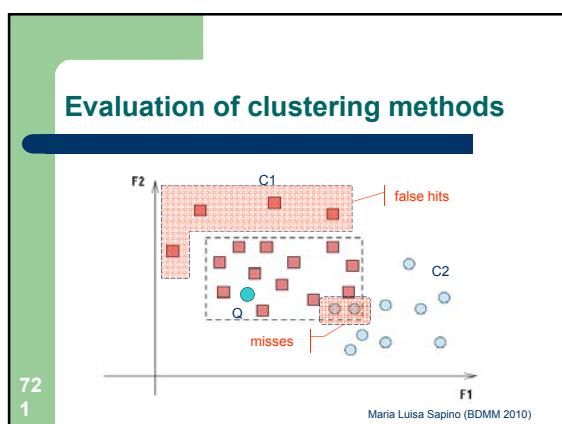
abbiamo query Q e vogliamo gli oggetti più simili alla query



se confrontiamo Q con i centroidi (rappresentanti)
allora restituiremo C1



gli oggetti nel rettangolo tratteggiato sono gli oggetti
che l'utente si aspettava di ricevere, quindi i risultati
ideali a fronte della query Q (il ground truth)



rispetto al risultato ideale ci sono FALSE HITS (oggetti
restituiti che non sono rilevanti per l'utente) e MISSES
(oggetti rilevanti per l'utente che non sono stati
restituiti)

Precision

- Precision

Retrieved and Relevant
Retrieved
measures the effect of false hits

Retrieved
Retrieved and Relevant
measures the effect of false hits

72
2

Maria Luisa Sapino (BDMM 2010)

DEFINIAMO DUE MISURE PER LE PRESTAZIONI DEL SISTEMA DI INFORMATION RETRIEVAL:

1) la PRECISIONE misura in quale percentuale gli oggetti restituiti sono anche rilevanti -> è tanto più alta quanti meno sono i false hits:
se nel lucido precedente non avessi i false hits in quel cluster, avrei una precisione del 100%

Precision and recall

- Precision
- Recall

Retrieved and Relevant
Retrieved
measures the effect of false hits

Retrieved and Relevant
Relevant
measures the effect of misses

72
3

Maria Luisa Sapino (BDMM 2010)

2) la RECALL misura in quale percentuale gli oggetti rilevanti, cioè quelli che l'utente vorrebbe ricevere, vengono restituiti
nel lucido di prima avrei avuto recall 100% se avessi restituito anche i due pallini azzurri in basso a dx

Precision and recall

- Precision
- Recall

Retrieved and Relevant
Retrieved
measures the effect of false hits

Retrieved and Relevant
Relevant
measures the effect of misses

Both should be closer to 1!!!!

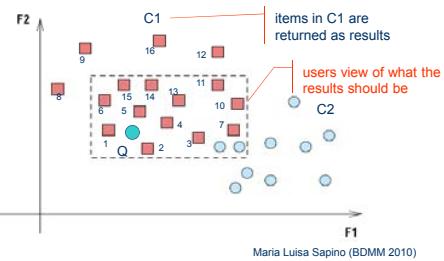
72
4

Maria Luisa Sapino (BDMM 2010)

idealmente sia P che R dovrebbero essere prossimi a 1, in modo tale da ricevere quasi tutto quello che interessa all'utente -> tipicamente però è difficile definire sistemi che restituiscano tutti e soli gli oggetti rilevanti!

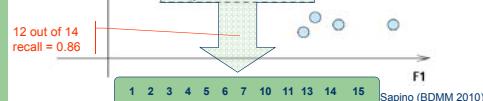
What if we also have rankings in the result???

72
5



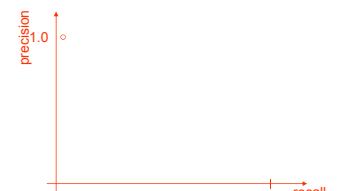
72
6

What if we also have rankings in the result???



72
7

What if we also have rankings in the result???



ora voglio tenere conto anche del RANK, ovvero dell'ordine in cui vengono restituiti i risultati: infatti l'utente vorrebbe per primi i risultati più rilevanti -> non mi basta che vengano restituiti tutti i risultati, ma voglio anche che mi vengano restituiti subito

come definire precisione e recall al passo K, ovvero dopo aver restituito K risultati?

(N.B. i risultati vengono restituiti nell'ordine crescente degli indici accanto ai quadratini rossi)

complessivamente qui ho recall = 12/14

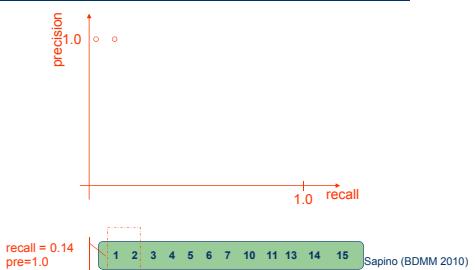
per tenere conto dell'ordine in cui vengono restituiti i risultati disegno una curva in questo modo:

1. al passo 1 ho restituito un risultato rilevante: precisione 1 e recall 1/14 (1 risultato su 14)
2. al passo 2: precisione 1 e recall 2/14
3. ecc... ecc..

8. al passo 8: precisione diminuisce e recall aumenta!

What if we also have rankings in the result???

72
8



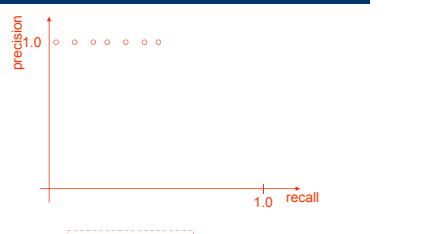
What if we also have rankings in the result???

72
9



What if we also have rankings in the result???

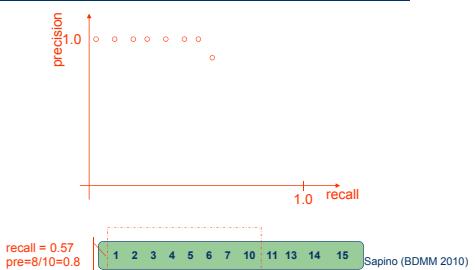
73
0



recall = 7/14
precisione = 1

What if we also have rankings in the result???

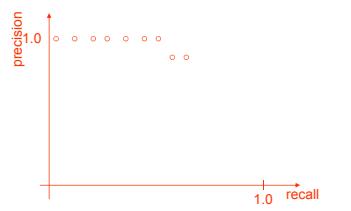
73
1



tipicamente ad un certo punto la precisione diminuisce mentre la recall aumenta gradualmente

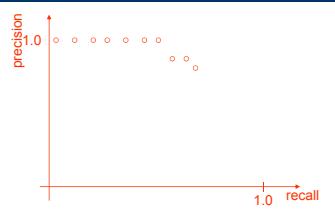
What if we also have rankings in the result???

73
2



What if we also have rankings in the result???

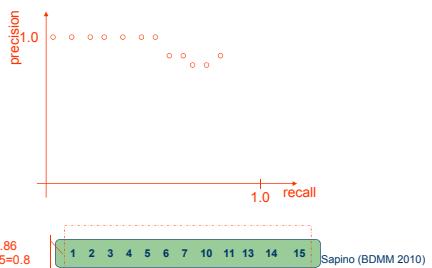
73
3



What if we also have rankings in the result???

73
4

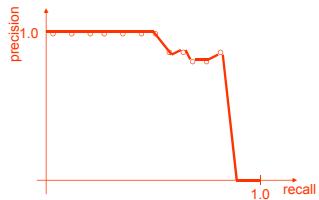
recall = 0.86
pre=12/15=0.8



Sapino (BDMM 2010)

Precision/recall curve

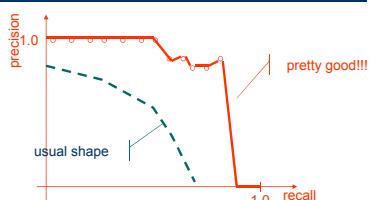
73
5



Sapino (BDMM 2010)

Precision/recall curve

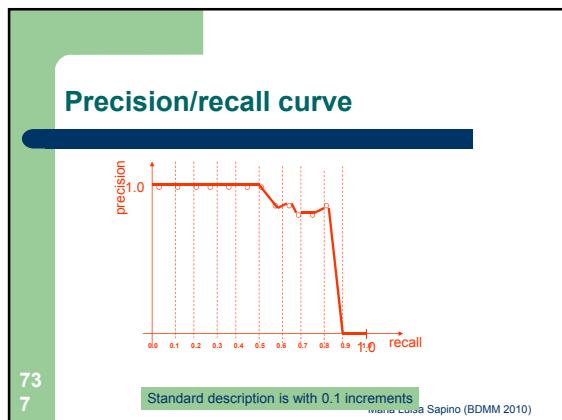
73
6



Maria Luisa Sapino (BDMM 2010)

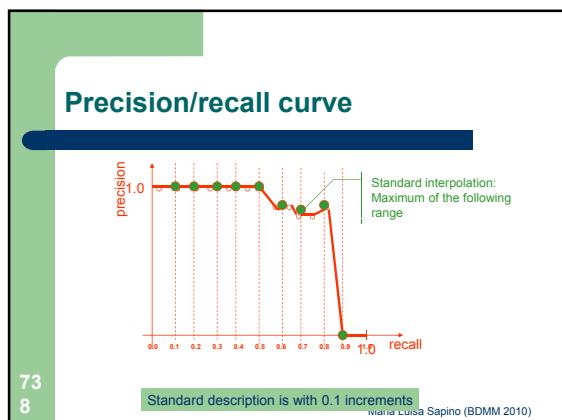
la forma tipica della curva PR è di questo tipo:
la recall aumenta ma la precisione diminuisce

si fa la descrizione standard per poter confrontare
sistemi diversi



73
7

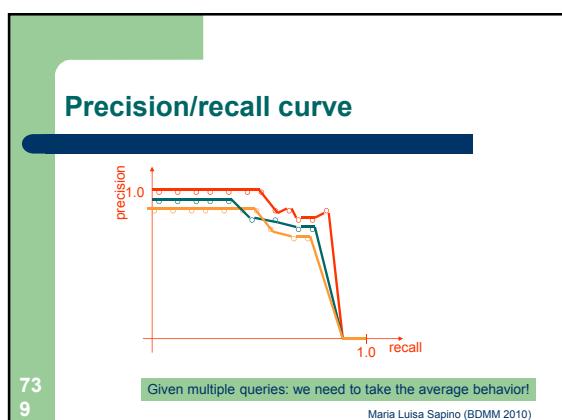
sull'asse delle ascisse si tengono intervalli di 0.1 per fare una descrizione standard



73
8

associamo dei valori alla curva in corrispondenza dei valori che ci interessano, quindi clusterizziamo rispetto a degli intervalli che abbiamo imposto sull'asse x

(come abbiamo fatto campionando i colori nell'istogramma nei colori)



73
9

quando su uno stesso sistema abbiamo una curva diversa per ogni query, allora prendiamo come rappresentante del sistema la curva media

Single-value summaries

- R-precision
 - # of relevant documents within first R
 - R is the total number of relevant documents in the result

74
0

Maria Luisa Sapino (BDMM 2010)

vogliamo tenere conto dell'ordine con cui sono restituiti gli oggetti

---> R-precision: quanti sono (in percentuale) gli oggetti rilevanti tra quelli che ho restituito fino al passo R?

quanto i primi R oggetti si discostano dal numero di oggetti ideali che mi aspetto vengano restituiti?

se uso come R la misura di TUTTI gli oggetti rilevanti, allora il valore della R-precision mi dà una misura globale del sistema (vorrei che i primi R risultati restituiti fossero proprio tutti e soli i rilevanti!)

Single-value summaries

- R-precision
 - # of relevant documents within first R
 - R is the total number of relevant documents in the result
- Example:
 - R = 14
 - # of relevant document in the first 14 is 11

1 2 3 4 5 6 7 10 11 13 14 15

- R-precision for this query is 11/14 = 0.876

74
1

Maria Luisa Sapino (BDMM 2010)

abbiamo visto due misure per valutare la qualità del sistema:

- 1) precisione
- 2) recall

idealmente dovrebbero essere alte entrambe, in realtà aumentando la recall tende a diminuire la precisione

Per avere una misura globale di un sistema (che tiene conto di precision e recall) ho bisogno di un unico numero in modo tale da poter confrontare agevolmente (la 'bontà' di) sistemi diversi.

Idea: calcolo la media aritmetica di precision e recall.

Ha il problema che un sistema con recall alto e precision bassa è visto globalmente come molto simile ad uno con recall basso e precision alta (la media è simile).

Anche un sistema con precision e recall con valori medi ha la stessa media dei due appena descritta.

Soluzione: usare la media armonica perché quest'ultima è alta solo quando tutti i valori sono alti. Quindi il sistema con precision e recall mediamente distribuito è migliore (ha media più alta) rispetto a quelli con precision alta e recall bassa o viceversa.

$$1/H = 1/2 * (1/P + 1/R)$$

quindi $H = 2PR / (P + R)$

se P e R sono entrambe alte anche H è alta (invece con media aritmetica non è detto!), quindi H è una buona misura della qualità del sistema

Harmonic mean

- The harmonic mean of n numbers (where $i = 1, \dots, n$) is

$$\frac{1}{H} = \frac{1}{n} \sum_{i=1}^n \frac{1}{x_i}$$

- Therefore, harmonic mean of $x = P$ and $y = R$

$$H(P,R) = 2PR / (P+R)$$

74
2

Maria Luisa Sapino (BDMM 2010)

Harmonic mean

- The harmonic mean of n numbers (where $i = 1, \dots, n$) is
$$\frac{1}{H} = \frac{1}{n} \sum_{i=1}^n \frac{1}{R_i},$$
- Therefore, harmonic mean of $x = P$ and $y = R$

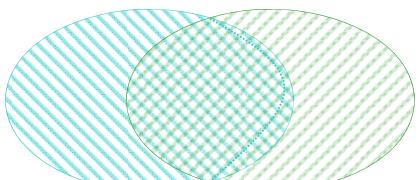
$$H(P,R) = 2PR / (P+R)$$

High only when both P and R are high
Maria Luisa Sapino (BDMM 2010)

74
3

Inoltre si possono associare dei pesi a precision e recall in modo tale da dare più o meno importanza ad uno o all'altro.

Coverage and Novelty



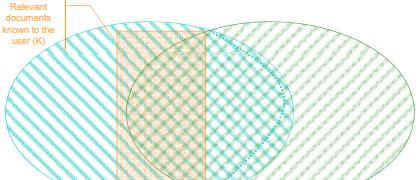
74
6

Relevant(rel) Returned(ret)
Maria Luisa Sapino (BDMM 2010)

altre misure della qualità del sistema che servono in un contesto diverso: ho una query che a fronte dei rilevanti (REL) restituisce un insieme di risultati (RET)

supponiamo di trovare un modo per raffinare la query per sovrapporre meglio l'insieme dei risultati ottenuti con quelli rilevanti

Coverage and Novelty

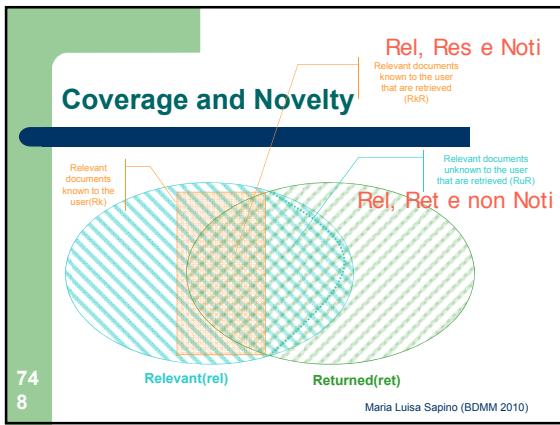


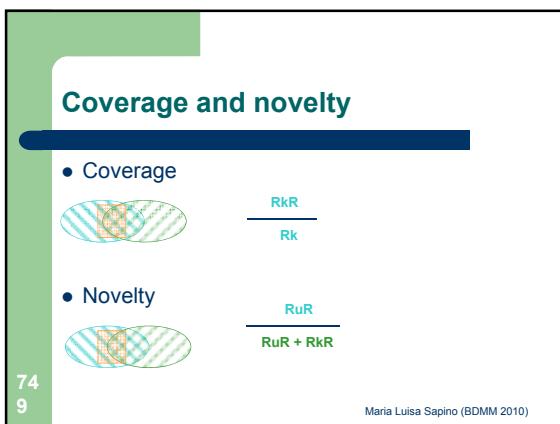
74
7

Relevant documents known to the user (K)
Relevant(rel) Returned(ret)
Maria Luisa Sapino (BDMM 2010)

* COPERTURA misura quanti degli oggetti che erano rilevanti e che erano presenti al passaggio precedente, sono ancora presenti nella query raffinata? -> voglio mantenere gli oggetti rilevanti!

* NOVITA' misura la percentuale degli oggetti interessanti che non avevo ancora visto e che sono presenti nella query raffinata



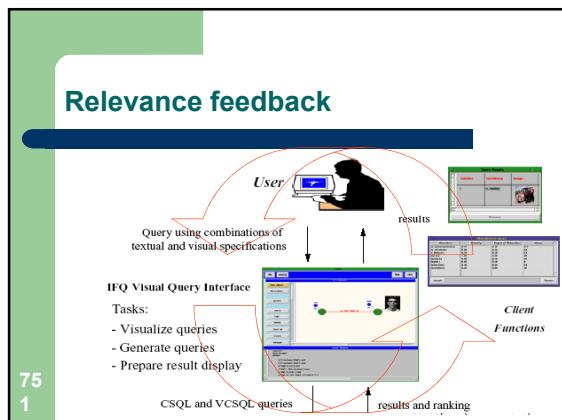


COPERTURA:
se mi elisse_dx nel rettangolo / rettangolo, quindi elementi Rel Ret e Noti / insieme dei Rel e Noti
--> voglio mantenere quelli rilevanti che avevo già restituito

NOVITA':
elementi Rel Ret e non Noti / elementi Rel e Ret (sia noti che non)

ha senso raffinare la query solo se mi aggiunge elementi nuovi e rilevanti: infatti voglio raffinare la query per trovare tutti e soli gli elementi che mi interessano, però non voglio perdere informazioni che avrei voluto conservare!

quando non c'è più novelty si può dire di aver approssimato sufficientemente la query e quindi non ha più senso raffinare



Relevance feedback

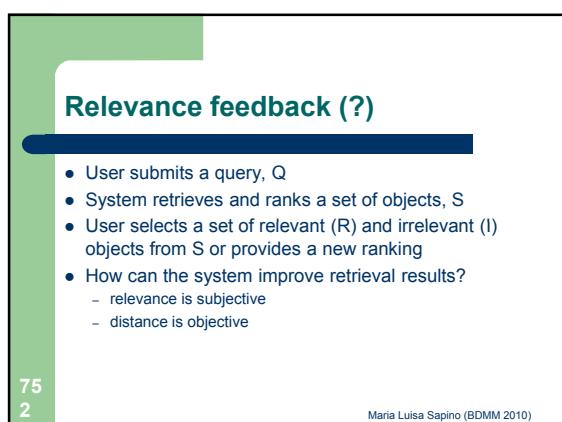
L'utente forma la query (input)

Il sistema mostra i risultati (output)

L'utente riceve questi risultati e fornisce ulteriori input al sistema in cui valuta i risultati ottenuti (input)

Per farlo contrassegna come 'soddisfacente' alcuni risultati e come 'insoddisfacenti' altri risultati.

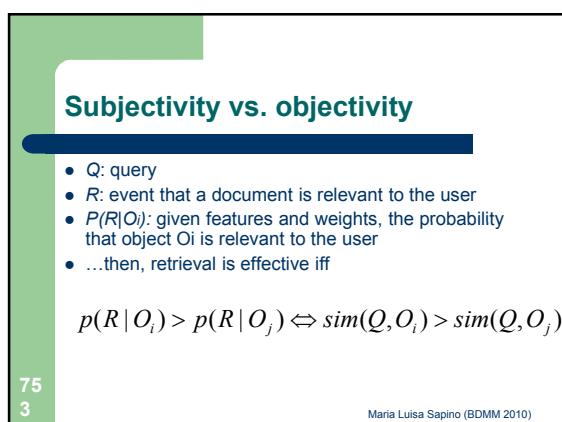
Il sistema apprende da questi input le RAGIONI che portano all'utente a fare certe scelte (di preferenze). Il sistema raffina allora la query basandosi su questi criteri.



tipicamente

- 1) l'utente fa una query
- 2) il sistema restituisce un insieme di oggetti S in un certo ordine
- 3) l'utente seleziona dalla risposta del sistema un insieme di oggetti rilevanti R e un insieme di oggetti irrilevanti I (che sono complementari), e può anche modificare il ranking

il sistema apprende dal feedback utente e lo usa per migliorare la qualità del retrieval: bisogna quindi fare convergere il feedback dell'utente (soggettivo) e la somiglianza tra i dati (oggettiva), perché il sistema deve basarsi sulla distanza oggettiva! Quindi la distanza tra due oggetti deve tenere conto anche da ciò che è stato specificato dall'utente.

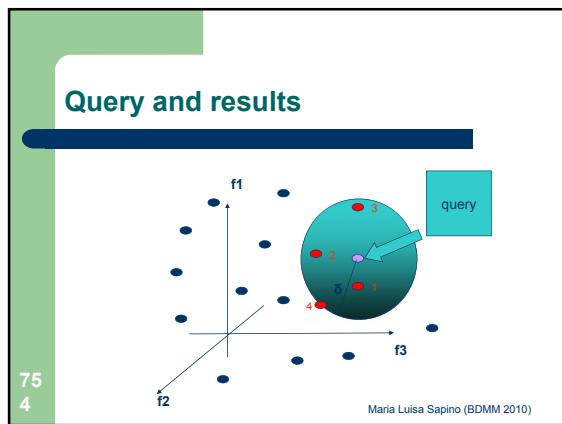


correliamo la dimensione soggettiva con quella oggettiva:

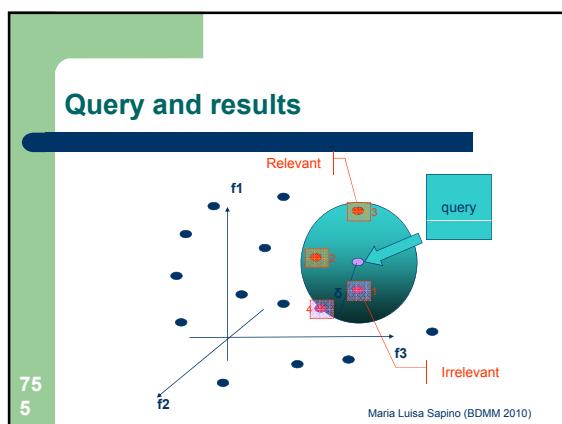
$$p(R|O_i) = p \text{ che l'oggetto } O_i \text{ se restituito dal sistema sia riconosciuto come rilevante dall'utente}$$

$$p(R|O_j) = p \text{ che l'oggetto } O_j \text{ se restituito dal sistema sia riconosciuto come rilevante dall'utente}$$

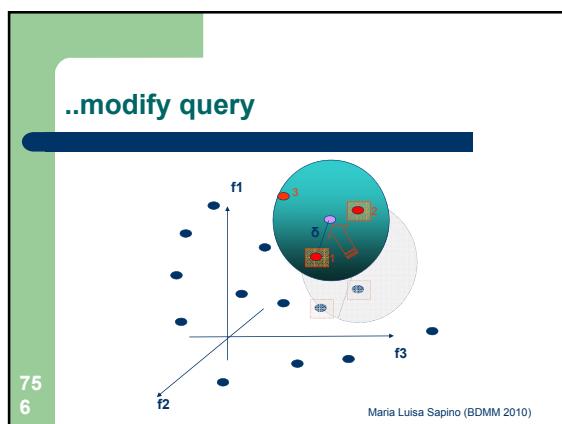
---> allora O_i è più rilevante di O_j se e solo se O_i è più simile OGGETTIVAMENTE alla query (rispetto alla metrica che uso per la misurazione) --> questo è l'obiettivo che il sistema deve raggiungere! Quindi più è probabile che un oggetto sia rilevante per c'utente maggior deve essere la sua distanza da una query qualsiasi.



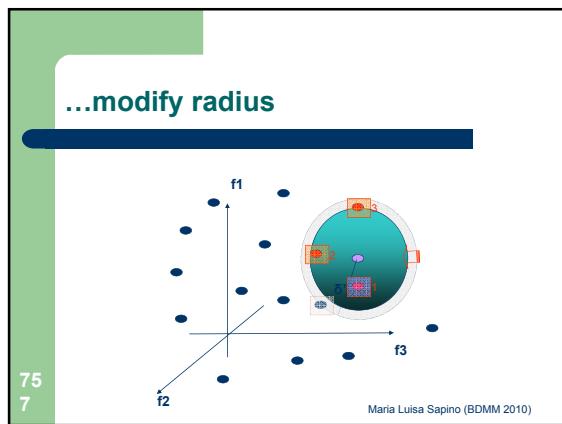
es. query di range che restituisce 4 oggetti



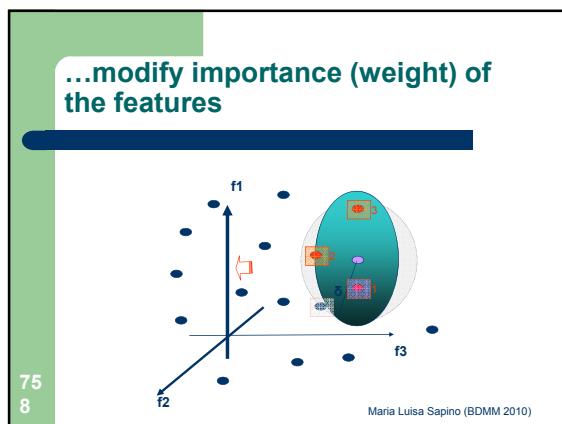
supponiamo che l'utente segni come rilevanti solo 2 e 3



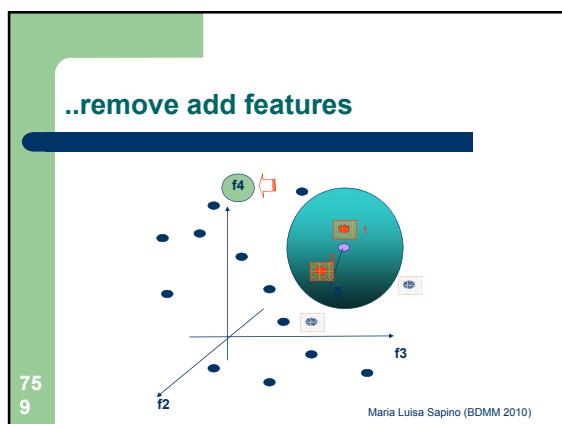
1) primo metodo per raffinare la query: correggere la posizione della query nello spazio (in questo caso spostarla verso l'alto) per "centrare" i risultati rilevanti



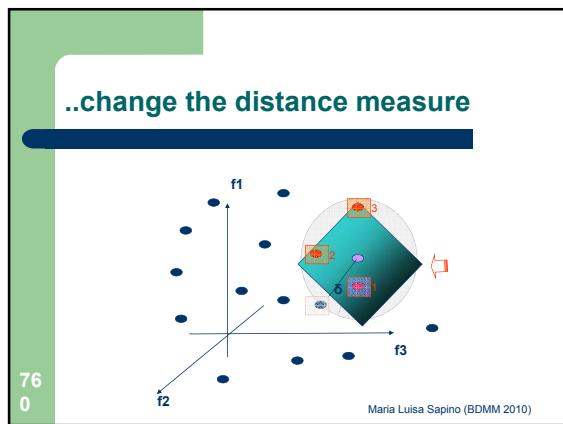
2) altra soluzione: correggere il range in cui cercare i valori



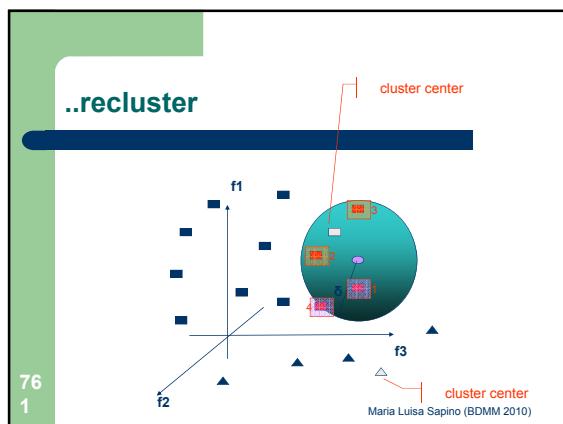
3) altra soluzione: agire sull'importanza delle features dando peso diverso a features diverse (ovvero dare importanza diversa alle diverse dimensioni) --> definisco una metrica in cui uso un range che non è più una sfera



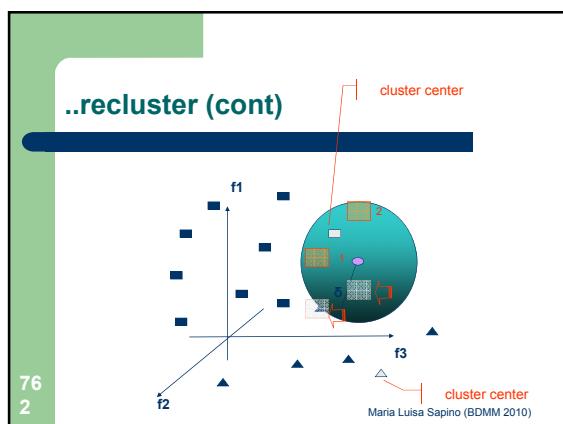
4) altra soluzione: lasciare cadere alcune features poco informative

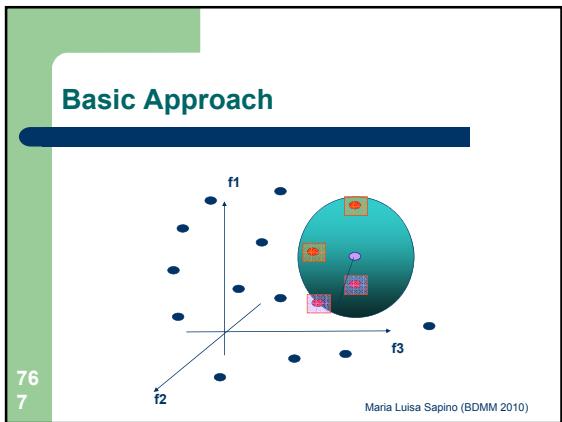


5) altra soluzione: cambiare la metrica (es: distanza di Manhattan)

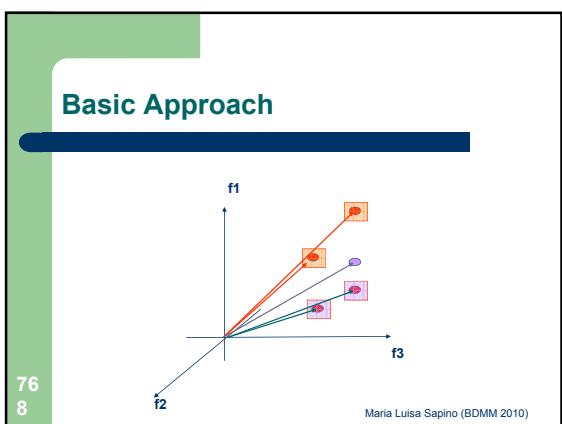


6) altra soluzione: se ho clusterizzato posso ri-clusterizzare cambiando alcuni parametri, ad es. # di cluster o dimensione massima di ciascun cluster

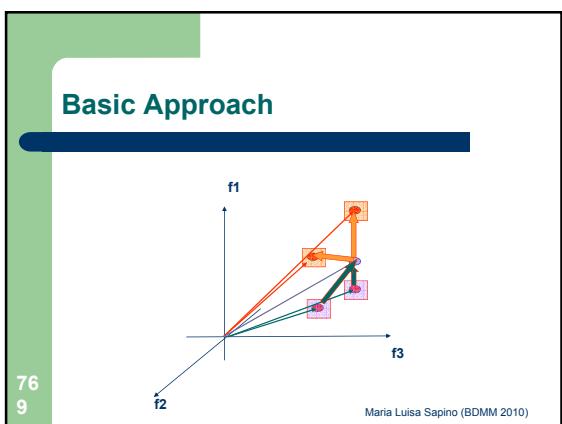




76
7



76
8



76
9

- * frecce arancioni: distanza della query dai rilevanti
- * frecce verdi: distanza della query dagli irrilevanti
(hanno segno opposto perché poi sommo)

Basic Approach

Maria Luisa Sapino (BDMM 2010)

Basic Approach

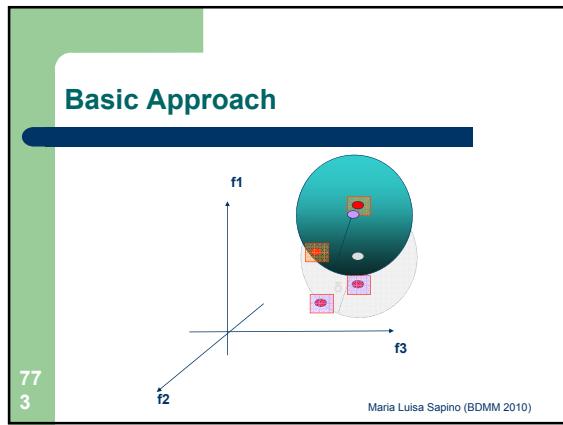
Maria Luisa Sapino (BDMM 2010)

Basic Approach

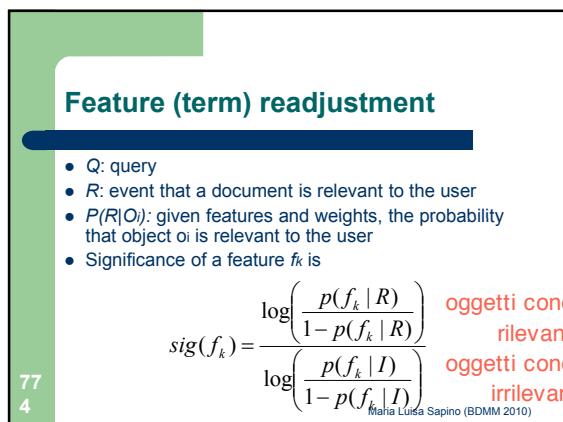
Maria Luisa Sapino (BDMM 2010)

la somma di tutti i vettori mi dà la direzione del vettore
rispetto a cui vado a spostare la query

freccia grigia mi dà la direzione, freccia rossa mi dice in
che punto fermarmi sulla direzione (è dato dai fattori c
che ho inserito nella formula per spostare la query!)

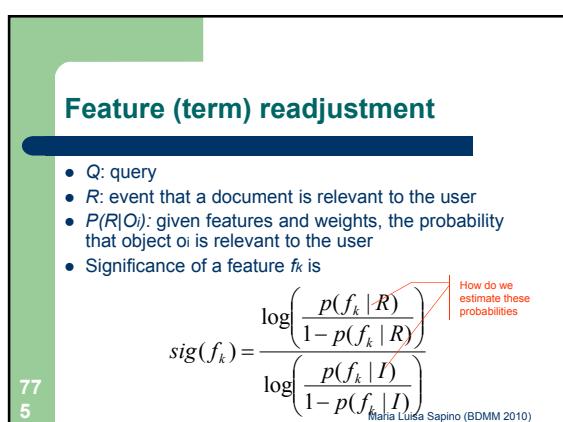


77
3



77
4

se la significatività è bassa allora la feature non porta informazione, ovvero non è sufficientemente discriminante



77
5

Nei casi in cui i due gruppi di oggetti (rilevanti, irrilevanti) non sono omogenei, quindi gli oggetti rilevanti sono vicini a quelli irrilevanti allora l'approccio basico non funziona bene perché in qualunque posizione io mi sposti sarò sempre vicino a oggetti rilevanti ma anche ad oggetti irrilevanti (es. oggetti rilevanti e irrilevanti sono disposti a cerchio attorno alla query).

Soluzione -> riaggiustamento dei termini ossia scelta delle feature da considerare più o meno importanti ai fini della valutazione della query.

- pesare feature in modo diverso
- lasciare cadere (non considerare) features
- aggiungere features

Per fare ciò occorre saper valutare quanto una feature è significativa (a feature significativa si da ad esempio peso minore oppure si lasciano cade ecc.)

$p(f_k | R)$ probabilità che un oggetto rilevante soddisfi la feature f_k .

$p(f_k | I)$ stessa cosa ma l'oggetto deve essere irrilevante.

Frazione a numeratore

-> numeratore: probabilità che la feature sia importante per un oggetto rilevante

-> denominatore: probabilità che la feature non sia importante per un oggetto rilevante

risultato -> importanza della feature per gli oggetti rilevanti

Frazione a denominatore -> stessa cosa per oggetti irrilevanti

-> $sig(f)$ è più alto tanto più la feature f è importante negli oggetti rilevanti e non importante in quelli irrilevanti (è caratteristica degli oggetti rilevanti, quindi l'utente anche se non l'ha specificato nella query desidera considerare tale feature).

Feature (term) readjustment

- Case I
 - If f_k is not a query term (not used in retrieval)

$$p(f_k | R) = p(f_k | Retrieved \& Relevant)$$

77
6

Maria Luisa Sapino (BDMM 2010)

Feature (term) readjustment

- Case II
 - If f_k is a query term (used in retrieval)

$$p(f_k | R) \neq p(f_k | Retrieved \& Relevant)$$

- There would be bias
 - Most retrieved objects will have f_k

77
7

Maria Luisa Sapino (BDMM 2010)

Feature (term) readjustment

- Let us assume binary feature and query
 - $o = \langle f_1, f_2, \dots, f_n \rangle \quad f_i = 0 \text{ or } 1$
 - $q = \langle w_1, w_2, \dots, w_n \rangle \quad w_i = 0 \text{ or } 1$
- Let us assume dot product as the similarity

$$sim(o, q) = \sum_{i=1}^n w_i f_i$$

77
8

Maria Luisa Sapino (BDMM 2010)

Come calcolare la probabilità $p(f_k | R)$?

Caso 1: se la feature non c'era nell'interrogazione significa che la probabilità di un oggetto di contenere tale feature è uniforme (è come scegliere casualmente oggetti che hanno o non hanno tale probabilità) o comunque dipende dalla probabilità di trovare oggetti con tale feature nell'intero database.

calcolo il rapporto:

oggetti rilevanti che contengono la feature f_k

oggetti rilevanti

Lo stesso vale per gli oggetti irrilevanti

2) caso delle features che erano presenti nella query: in questo caso i restituiti e rilevanti sono condizionati dalla query che ho fatto, quindi non posso calcolare p come prima
---> non c'è più indipendenza tra l'insieme degli oggetti restituiti e la feature k !

----> devo fare qualche manipolazione per riportarmi ad un caso di indipendenza tra feature presente nel database e query.

* consideriamo feature e query binarie: può avere solo valore 0 o 1

La query ha peso 0 o 1 per ogni dimensione per denotare che la corrispondente feature interessa o meno nella query.

Assumiamo anche che la somiglianza tra due oggetti sia definita tramite il dot product di vettori.

Conta il numero di feature richieste dalla query soddisfatte dall'oggetto f_i .

Es. se $o = \langle 1, 0, 0, 1 \rangle$ e $q = \langle 1, 0, 0, 0 \rangle$ l'oggetto o ha similarità 1 perché soddisfa una feature richiesta dalla query, mentre $o' = \langle 1, 1, 1, 1 \rangle$ ha similarità 2 perché

Feature (term) readjustment

- If a document is returned, then

$$sim(o, q) = \sum_{i=1}^n w_i f_i > T$$

77
9

Maria Luisa Sapino (BDMM 2010)

vogliamo che gli oggetti restituiti assomiglino almeno T alla query (= hanno almeno T features corrispondenti a 1)

Feature (term) readjustment

- Let's focus on a specific feature

$$sim(o, q) = w_j f_j + \sum_{i \in \{1..n\} \setminus \{j\}} w_i f_i > T$$

or

$$sim(o, q) = w_j f_j + sim_{(-j)}(o, q) > T$$

somiglianza escludendo la feature j

78
0

Maria Luisa Sapino (BDMM 2010)

siamo interessati ad una specifica feature j, quindi la mettiamo in evidenza:

w_j f_j è il singolo contributo della feature j che abbiamo estratto dalla sommatoria, ma allora

sim_{-j}(O, Q) è la somiglianza tra O e Q escludendo la feature j

Feature (term) readjustment

$$sim(o, q) = w_j f_j + sim_{(-j)}(o, q) > T$$

	f _j =1	f _j =0
sim_{(-j)}(o, q) ≤ T	a	0
sim_{(-j)}(o, q) > T	b	c

78
1

$$|relevant \& retrieved| = a + b + c$$

Maria Luisa Sapino (BDMM 2010)

tavola che esprime la cardinalità dei diversi insiemi degli oggetti restituiti

N.B. ho 0 nella cella [1, 2] perché non ho nessun oggetto restituito che non sarebbe stato restituito senza la feature j e non soddisfa la feature j, altrimenti non lo avrei tra i restituiti!

considero come varia l'insieme dei risultati con o senza una specifica feature:

prima riga -> in assenza della feature j l'oggetto non sarebbe stato restituito

seconda riga -> anche senza la feature j l'oggetto sarebbe stato restituito (INDIPENDENTE DA J!!!!)

prima colonna -> quanti sono gli oggetti, tra quelli non restituiti senza feature j, che soddisfano la feature j

seconda colonna -> quanti sono gli oggetti, tra quelli non restituiti senza feature j, che non soddisfano la feature j = 0

* ho a elementi per cui j è stato fondamentale per il retrieval

* ho b elementi tra quelli restituiti che sarebbero restituiti cmq anche senza j e per cui j vale 1

* ho b elementi tra quelli restituiti che sarebbero restituiti cmq anche senza j e per cui j vale 0

--> I Rel & Ret I = a + b + c

10

Feature (term) readjustment

$$p(f_k | \text{Ret} \& \text{Rel}) = \frac{p((f_k = 1) \wedge (\text{sim}_{(-k)}(o, q) > T))}{p(\text{sim}_{(-k)}(o, q) > T)}$$

Without bias of f_k

ho escluso gli elementi nella seconda riga della tabella precedente

Maria Luisa Sapino (BDMM 2010)

78
2

esprimiamo la probabilità che una certa feature sia presente nell'insieme Ret&Rel (sui cui vogliamo contare le frequenze)

* ho escluso la seconda riga perché i risultati erano indipendenti dalla presenza di f_k

P che ci sia f_k e che sia nella seconda riga (infatti sim supera T anche senza la feature k), il tutto diviso per P che gli oggetti mi verrebbero restituiti anche senza considerare la feature k

----> b / (b+c)

Feature (term) readjustment

$$p(f_k | \text{Ret} \& \text{Rel}) = \frac{p((f_k = 1) \wedge (\text{sim}_{(-k)}(o, q) > T))}{p(\text{sim}_{(-k)}(o, q) > T)}$$

independent

↓

$$p(f_k | \text{Ret} \& \text{Rel}) = p(f_k = 1)$$

Maria Luisa Sapino (BDMM 2010)

78
3

Ma i due eventi della probabilità al numeratore sono indipendenti (prob che $f_k=1$ e che $\text{sim}_{(-k)}(o, q) > T$) per cui per le leggi della probabilità si può fare il prodotto

$$p(f_k=1) * p(\text{sim}(o, q) > T)$$

per cui si può semplificare a denominatore ottenendo

$$p(f_k | \text{Ret} \& \text{Rel}) = p(f_k = 1)$$

Feature (term) readjustment

$$p(f_k | \text{Ret} \& \text{Rel}) = \frac{p((f_k = 1) \wedge (\text{sim}_{(-k)}(o, q) > T))}{p(\text{sim}_{(-k)}(o, q) > T)}$$

independent

↓

$$p(f_k | \text{Ret} \& \text{Rel}) = p(f_k = 1) = p(f_k = 1 | \text{sim}_{(-k)}(o, q) > T)$$

Maria Luisa Sapino (BDMM 2010)

78
4

Ma calcolare $p(f_k = 1)$ significa esaminare l'intero database e effettuare il rapporto oggetti con $f_k=1$ / oggetti nel db. Ma siccome gli eventi $f_k=1$ e $\text{sim}(o, q) > T$ sono indipendenti, significa che posso esaminare solo lo spazio di oggetti restituiti tali che $\text{sim}(o, q) > T$ tramite

$$p(f_k = 1 | \text{sim}(o, q) > T)$$

quindi tale probabilità si calcola come il rapporto

numero di oggetti rilevanti con $f_k=1$ / numero di oggetti rilevanti

Feature (term) readjustment

$$p(f_k \mid \text{Ret} \& \text{Rel}) = \frac{p((f_k = 1) \wedge (\text{sim}_{(-k)}(o, q) > T))}{p(\text{sim}_{(-k)}(o, q) > T)}$$



$$p(f_k \mid R) = \frac{b}{b + c}$$

Maria Luisa Sapino (BDMM 2010)

78
6

sapendo che il sistema ci restituisce gli oggetti ordinati,
vorremmo che la somiglianza tra gli oggetti i e j fosse
coerente con la rilevanza dei corrispondenti oggetti per
l'utente: deve avere più P di essere rilevante l'oggetto
più simile alla query

il sistema può solo agire sulla somiglianza oggettiva tra
gli oggetti, non può agire sulla rilevanza

Ranking

- Q: query
- R: event that a document is relevant to the user
- $P(R|O_i)$: given features and weights, the probability that object o_i is relevant to the user
- ...then, retrieval is effective iff

$$p(R|O_i) > p(R|O_j) \Leftrightarrow \text{sim}(Q, O_i) > \text{sim}(Q, O_j)$$

78
7

Maria Luisa Sapino (BDMM 2010)

riscriviamo la parte sinistra applicando il teorema di
Bayes che regola le probabilità condizionate

Ranking

$$p(R|O_i) > p(R|O_j) \Leftrightarrow \text{sim}(Q, O_i) > \text{sim}(Q, O_j)$$

- Let's try to rewrite the first half of the equation using Bayes theorem

78
8

Maria Luisa Sapino (BDMM 2010)

Bayes Theorem

$$p(A|B) = \frac{p(B|A)p(A)}{p(B)}$$

78
9

Maria Luisa Sapino (BDMM 2010)

P che l'oggetto O sia riconosciuto come rilevante
espressa come P che l'oggetto riconosciuto come
rilevante sia proprio l'oggetto O

Bayes Theorem

$$p(A|B) = \frac{p(B|A)p(A)}{p(B)}$$

$$p(A|B) = \frac{p(B|A)p(A)}{p(B|A)p(A) + p(B|\neg A)p(\neg A)}$$

79
0

Maria Luisa Sapino (BDMM 2010)

Relevance of two objects

$$p(R|O_i) = \frac{p(O_i|R)p(R)}{p(O_i|R)p(R) + p(O_i|I)p(I)}$$

>

$$p(R|O_j) = \frac{p(O_j|R)p(R)}{p(O_j|R)p(R) + p(O_j|I)p(I)}$$

79
1

Maria Luisa Sapino (BDMM 2010)

P che O_i sia rilevante
* not A = not R = I

vogliamo che valga questa diseguaglianza

Relevance of two objects

$$\frac{p(O_i | R)}{p(O_i | I)} > \frac{p(O_j | R)}{p(O_j | I)}$$

79
3

Maria Luisa Sapino (BDMM 2010)

applicando un po' di manipolazioni algebriche (è sufficiente invertire numeratore/denominatore da entrambe le parti e invertire il segno) si ottiene questa disegualanza: questa è una

rappresentazione più semplice in cui ho tutte le P condizionate in cui l'evento noto è l'essere rilevante/irrilevante e devo stimare la P che venga restituito un oggetto piuttosto che un altro

...so we have

$$\frac{p(O_i | R)}{p(O_i | I)} > \frac{p(O_j | R)}{p(O_j | I)} \Leftrightarrow sim(Q, O_i) > sim(Q, O_j)$$

How do we compute these probabilities???

79
4

Maria Luisa Sapino (BDMM 2010)

come calcoliamo questa P?

...so we have

- Let us assume features are independent
 - $o = \langle f_1, f_2, \dots, f_n \rangle$
 - $q = \langle w_1, w_2, \dots, w_n \rangle$

$$p(O_i | R) = \prod_{k=1}^n p(f_{i,k} | R) \quad p(O_i | I) = \prod_{k=1}^n p(f_{i,k} | I)$$

79
5

Maria Luisa Sapino (BDMM 2010)

assumiamo che le features siano tra loro indipendenti
---> così la P che un oggetto rilevante sia proprio O_i è data dal prodotto delle P che un oggetto rilevante soddisfi tutte le features dell'oggetto i (un oggetto è una congiunzione di features)

dobbiamo capire come aggiustare i pesi in modo tale che il risultato sia quello che vogliamo

es: se cerco foto di donne brune e O_i è una foto di una donna con i capelli bruni, allora la P che un oggetto rilevante sia proprio O_i è data dalla P che un oggetto rilevante abbia entrambe le features (sesso e colore dei capelli sono features indipendenti)

...so we have

- Let us assume features are independent

$$p(O_i | R) = \prod_{k=1}^n p(f_{i,k} | R) \quad p(O_i | I) = \prod_{k=1}^n p(f_{i,k} | I)$$

$$\frac{p(O_i | R)}{p(O_i | I)} > \frac{p(O_j | R)}{p(O_j | I)} \Leftrightarrow \frac{\prod_{k=1}^n p(f_{i,k} | R)}{\prod_{k=1}^n p(f_{i,k} | I)} > \frac{\prod_{k=1}^n p(f_{j,k} | R)}{\prod_{k=1}^n p(f_{j,k} | I)}$$

Maria Luisa Sapino (BDMM 2010)

79
6

- siccome ho produttorie sugli stessi indici posso portare fuori la produttoria e applicarla al rapporto

- applico il log ai due prodotti:

- uso proprietà che il log di una produttoria è uguale alla sommatoria dei logaritmi

---> questo mi serve per arrivare alla misura della somiglianza: infatti il DOT PRODUCT fa la sommatoria dei prodotti

...so we have

- Let us assume features are independent

$$p(O_i | R) = \prod_{k=1}^n p(f_{i,k} | R) \quad p(O_i | I) = \prod_{k=1}^n p(f_{i,k} | I)$$

$$\frac{p(O_i | R)}{p(O_i | I)} > \frac{p(O_j | R)}{p(O_j | I)} \Leftrightarrow \sum_{k=1}^n \log \frac{p(f_{i,k} | R)}{p(f_{i,k} | I)} > \sum_{k=1}^n \log \frac{p(f_{j,k} | R)}{p(f_{j,k} | I)}$$

Maria Luisa Sapino (BDMM 2010)

79
7

- se uso DOT PRODUCT come misura di somiglianza

- se il peso della feature k-esima dell'oggetto i-esimo

è dato dal log del rapporto $p(f_{ik} | R) / (f_{ik} | I)$

- se uso come query $\langle 1, 1, \dots, 1 \rangle$

allora $\sum \log p(f_{ik} | R) / p(f_{ik} | I)$ è esattamente la somiglianza tra l'oggetto O e la query Q:

infatti è dato da $\sum \log p(f_{ik} | R) / p(f_{ik} | I) * 1$

ho costruito i pesi in modo tale da garantire che la disugualanza valga

...so we have

- Let us assume features are independent
 - use dot product as the similarity measure
 - use $\log \frac{p(f_{ik} | R)}{p(f_{ik} | I)}$ as the weight of the kth feature
 - use $\langle 1, 1, \dots, 1 \rangle$ as the query!!!

$$\frac{p(O_i | R)}{p(O_i | I)} > \frac{p(O_j | R)}{p(O_j | I)} \Leftrightarrow \sum_{k=1}^n \log \frac{p(f_{i,k} | R)}{p(f_{i,k} | I)} > \sum_{k=1}^n \log \frac{p(f_{j,k} | R)}{p(f_{j,k} | I)}$$

Maria Luisa Sapino (BDMM 2010)

79
8

...what if features are not independent?

- Let us assume features are not independent
 - $o = \langle f_1, f_2, \dots, f_n \rangle$
 - $q = \langle w_1, w_2, \dots, w_n \rangle$

$$p(O_i | R) \neq \prod_{k=1}^n p(f_{i,k} | R) \quad p(O_i | I) \neq \prod_{k=1}^n p(f_{i,k} | I)$$

79
9

Maria Luisa Sapino (BDMM 2010)

cosa succede se le features non sono indipendenti?

allora non è più vero che $p(O_i | R)$ è uguale alla produttoria di $p(f_{ik} | R)$ e che $p(O_i | I)$ è uguale alla produttoria di $p(f_{ik} | I)$!

...what if features are not independent?

- Let us assume features are not independent
 - $o = \langle f_1, f_2, \dots, f_n \rangle$
 - $q = \langle w_1, w_2, \dots, w_n \rangle$
- How can we incorporate term dependence???

80
0

Maria Luisa Sapino (BDMM 2010)

come tenere conto del fatto che le features non sono indipendenti?

...so we have

- Let us assume features are not independent
 - $o = \langle f_1, f_2, \dots, f_n \rangle$
 - $q = \langle w_1, w_2, \dots, w_n \rangle$
- How can we incorporate term dependence???
- Degree of approximation.....

$$I(p1, p2) = \sum_x p1(x) \log \frac{p1(x)}{p2(x)}$$

- $p1 = p2$ implies that $I = 0$
- $p1 \neq p2$ implies that $I > 0$

80
1

Maria Luisa Sapino (BDMM 2010)

misuriamo il grado di indipendenza tra due distribuzioni di probabilità $p1$ e $p2$:

- se $p1 = p2$, allora $I = 0$ e quindi $I = 0$
 - se $p1 \neq p2$, allora $I > 0$ e I è tanto maggiore quanto più $p1$ e $p2$ sono diversi

...so we have

- Degree of approximation.....

$$I(p_1, p_2) = \sum_x p_1(x) \log \frac{p_1(x)}{p_2(x)}$$

- $p_1=p_2$ implies that $I=0$; $p_1 > p_2$ implies that $I>0$

- Degree of dependence between f_i and f_j

$$D_{ij} = I(p(f_i \wedge f_j), p(f_i)p(f_j))$$

80
2

If the two terms are independent, then D_{ij} will be 0!!! (SDMM 2010)

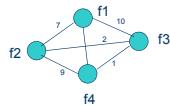
voglio capire quanto la feature f_i e la feature f_j siano tra loro indipendenti

- so che se f_i e f_j fossero completamente indipendenti, sarebbe vero che $P(f_i \wedge f_j) = P(f_i) * P(f_j)$ ---> uso questo criterio per giudicare quanto due features siano tra loro indipendenti, ovvero

due features sono tanto più indipendenti quanto più sono simili le P congiunte e il prodotto delle due P delle features

---> se le features sono indipendenti, allora $D_{ij} = 0$ altrimenti $D_{ij} > 0$ e il valore è tanto più alto quanto maggiore è la differenza tra $P(f_i \wedge f_j)$ e $P(f_i) * P(f_j)$

Dependence graph



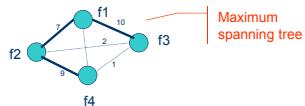
80
3

If the two terms are independent, then D_{ij} will be 0!!! (SDMM 2010)

1) costruisco un grafo delle dipendenza tra features: per ogni coppia di features, le due features sono unite da un arco che ha come peso il valore del loro grado di indipendenza

Si possono rimuovere gli archi con un grado di dipendenza basso (che collega nodi che possono considerarsi indipendenti).

Dependence graph



80
4

If the two terms are independent, then D_{ij} will be 0!!! (SDMM 2010)

2) calcolo Maximum Spanning Tree perché non posso gestire tutte l'informazione coppia per coppia e quindi mi concentro sulle features che sono più dipendenti

Dependence graph

```
graph TD; f1((f1)) -- 7 --> f2((f2)); f1 -- 10 --> f3((f3)); f2 -- 2 --> f3; f2 -- 9 --> f4((f4)); f3 -- 1 --> f4;
```

$$p(f_1 \wedge f_2 \wedge f_3 \wedge f_r) = p(f_1)p(f_2 | f_1)p(f_3 | f_1)p(f_4 | f_2)$$

If the two terms are independent, then D_{ij} will be 0!!!

© 2010 University of Washington - BDMM 2010

a questo punto posso vedere l'oggetto di partenza come:

$P(\text{radice del MST}) * \text{\prod } P(\text{figlio} \mid \text{padre})$

---> so calcolare queste P condizionate:
ad es. per $p(f_4 | f_2)$ basta guardare tra gli oggetti
restituiti quanti soddisfano f_4 e tra questi guardo
quanti soddisfano anche f_2

Dependence graph

$$p(f_1 \wedge f_2 \wedge f_3 \wedge f_r) = p(f_1)p(f_2 | f_1)p(f_3 | f_1)p(f_4 | f_2)$$

$p(O_i | R)$ can be computed using the distribution of the features in $R!!!$

$p(O_i | I)$ can be computed using the distribution of the features in $I!!!$

80

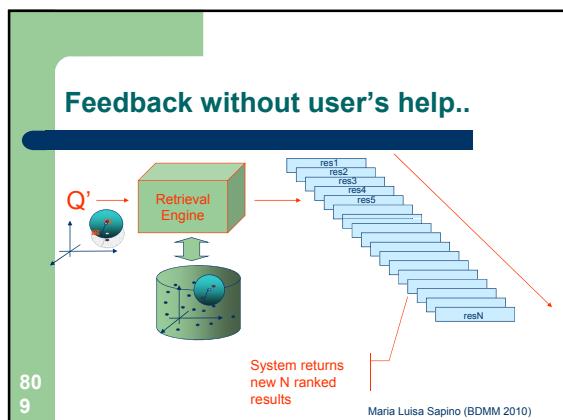
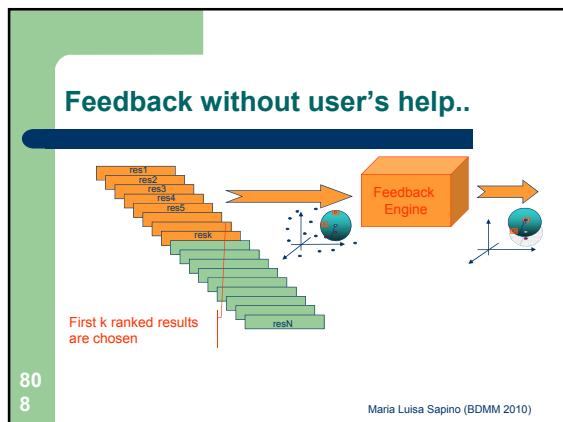
6

Maria Luisa Sapino (BDMM 2010)

come gestire feedback implicito o pseudo-feedback?

-l'utente non dà feedback, quindi ci basiamo solo sulle informazioni date dal sistema, che ci restituisce delle immagini ordinate in un certo modo

Feedback without user's help..



- assumiamo che il sistema abbia lavorato bene e che effettivamente i primi k risultati siano soddisfacenti per l'utente e che gli ultimi j risultati siano quelli che interessano meno all'utente

---> sulla base di questo risultato raffino la query

in pratica l'utente non seleziona i rilevanti, ma dice solo al sistema di riprovare la query

+) è meno impegnativo per l'utente

-) grosso peso della prima query: a fronte di una query inadeguata ovviamente si avranno anche risultati inadeguati

Multimedia query processing

- First major issue
 - imperfections (fuzziness)
 - Second issue
 - ranking
 - Third issue
 - expensive predicates (user defined functions)

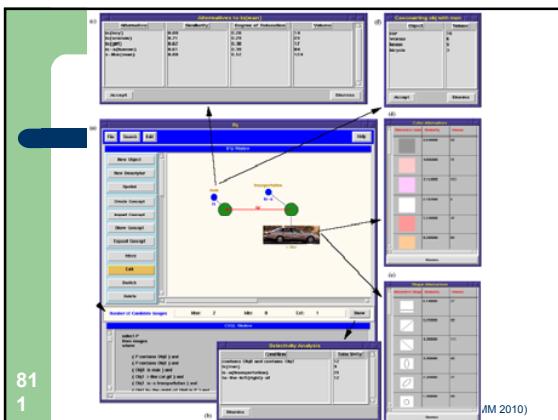
Maria Luisa Sapino (BDMM 2010)

Per gestire dati multimediali non possiamo utilizzare il modello relazionale perché dobbiamo modellare incertezza, imperfezione (es. matching parziale) ecc.

Ma il matching parziale si basa sul concetto di fuzziness (es. essere vestito di nero cosa significa, quando un colore cessa di essere nero? Quanto è nero? oppure quanto una persona è giovane? Esiste una vera e propria linea di confine tra essere giovane o vecchio?) e non sul concetto di confronto esatto.

Inoltre per il ranking occorre definire dei criteri di ordinamento globale che combinano varie caratteristiche locali (es. colore di un'immagine e luminosità vanno ordinate in modo tale da tenere conto di entrambi gli aspetti, es. immagine rossa e chiara è vicina ad una rossa e scura e anche una rosa e chiara)

Alcuni predicati (immagine che contiene un oggetto, che è simile a ...) sono molto costosi, per cui occorre effettuare ottimizzazione -> ad es. se un oggetto per essere selezionato da una query deve soddisfare due predicati, prima verifico che soddisfi quello meno costoso computazionalmente, e dopo quello più costoso, perché se non soddisfa il primo non soddisfa la query per cui è inutile fare un confronto che sarebbe molto costoso.



A multimedia query

```

select image P, object object1, object object2
where P contains object1
    and P contains object2
    and object1.semantical.property s-like "mountain"
    and object1.image.property image_match "Fuji_mountain.gif"
    and object2.semantical.property is "lake"
    and object2.image.property image_match "lake.image_sample.gif"
    and object1.position is_above object2.position

```

Maria Luisa Sapino (BDMM 2010)

A multimedia query

Query
Fuji Mountain
↓
Lake

Crisp

```
select image P, object object1, object object2
where P [contains] object1
and P [contains] object2
and object1.semantical.property [s_like] "mountain"
and object1.image.property [image_match] "Fuji.mountain.gif"
and object2.semantical.property [is] "lake"
and object2.image.property [image_match] "lake_image_sample.gif"
and object1.position [is_above] object2.position
```

Fuzzy (imperfect)

81
3

Maria Luisa Sapino (BDMM 2010)

- all'interno della query multimediale ci possono essere predicati
- CRISP: hanno solo risposta binaria, es. contenimento di un oggetto, presenza di un particolare colore
 - FUZZY: una componente può essere presente con un certo grado, es. predominanza di un certo colore
- Predicati
- P contains object: l'immagine P deve contenere l'oggetto 'object'. E' un predicato CRISP perché l'oggetto è stato rilevato (perché è stato identificato tramite tag o riconoscimento tramite algoritmo) oppure no.
 - A s_like B: A ha come proprietà semantica B.
 - A is B: A ha semantica B.
 - A image_match B: A è simile a B.
 - A is_above B: l'oggetto A è sopra l'oggetto B.

Query...and results...

Query
Fuji Mountain
↓
Lake

(a)
0.90
0.00
0.90
Fuji Mountain
Lake

(b)
0.5
1.0
0.5
Mountain
Lake

(c)
0.8
0.5
0.0
Fuji Mountain
Forest

(d)
1.0
0.5
0.5
Fuji Mountain
Lake

81
4

Maria Luisa Sapino (BDMM 2010)

ciascun predicato atomico (= non ulteriormente decomponibile) viene valutato in modo autonomo
rispetto agli altri e associato ad un grado di soddisfacimento: alla fine bisogna combinare questi risultati parziali per restituire il grado di soddisfacimento globale della query

Reasons for imperfection

- Similarity between features (yellow/orange)
- Imperfections in the feature extraction algorithms
- Imperfections in the query formulation methods
- Partial match requirements
- Imperfections in the index structures and clustering algorithms

81
5

Maria Luisa Sapino (BDMM 2010)

Fuzzy set..

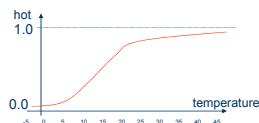
- Fuzzy set F with domain D is defined using a membership function $\mu_F : D \rightarrow [0, 1]$.
- A crisp (conventional) set C with domain D is defined using a membership function $\mu_C : D \rightarrow \{0, 1\}$.
- A fuzzy set corresponds to a fuzzy predicate

81
6

Maria Luisa Sapino (BDMM 2010)

insieme CRISP invece ha funzione di appartenenza che va dal dominio all'insieme binario $\{0, 1\}$, quindi si può vedere come caso particolare di insieme fuzzy dove l'unico caso possibile di valore è 0 (non appartenenza) e 1 (appartenenza)

Example



81
7

Maria Luisa Sapino (BDMM 2010)

Un insieme fuzzy F è una collezione di elementi definita da una funzione di appartenenza $\mu : D \rightarrow [0, 1]$ tale che

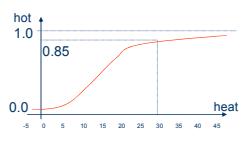
x in F con grado di appartenenza $\mu(x)$ e non appartiene a F se $\mu(x) = 0$.

E' diverso dagli insiemi tradizionali (insieme crisp) in cui un oggetto appartiene o non appartiene a tale insieme. Negli insiemi fuzzy un oggetto può appartenere più o meno (con più o meno confidenza) a tale insieme.

- rappresento insieme fuzzy "essere caldo"

- man mano che la temperatura sale aumenta il grado di appartenenza della temperatura all'insieme

Example



Hot(29°) = 0.85

81
8

Maria Luisa Sapino (BDMM 2010)

mappa valore del dominio al grado di soddisfacimento del predicato che caratterizza l'insieme.

Così come è possibile definire un predicato (crisp) su un insieme (crisp), è anche possibile fare lo stesso con gli insiemi fuzzy.

p predicato definito sull'insieme P

$p(x) = i$ SSE $x \in P$ con grado di appartenenza i

Empty fuzzy set

$$\forall x \in X : f_\phi(x) = 0$$

81
9

Maria Luisa Sapino (BDMM 2010)

L'insieme fuzzy vuoto è l'insieme definito dalla funzione f tale che $f(x) = 0$ per ogni x appartenente al dominio della funzione f .

Universal fuzzy set

$$\forall x \in X : f_u(x) = 1$$

82
0

Maria Luisa Sapino (BDMM 2010)

L'insieme fuzzy universale è definito tramite la funzione f tale che $f(x) = 1$ per ogni x appartenente al dominio della funzione f .

α -Universal fuzzy set

$$\forall x \in X : f_{X^{[\alpha]}}(x) = \alpha$$

82
1

Maria Luisa Sapino (BDMM 2010)

Un insieme fuzzy universale rispetto ad α è un insieme tale che $f(x) = \alpha$ per ogni x nel dominio di f .

Support of a fuzzy set, F

$$\text{supp}(F) = \text{def} \{x \in X \mid f_F(x) > 0\}$$

82
2

Maria Luisa Sapino (BDMM 2010)

Il supporto di un insieme fuzzy F è l'insieme degli elementi del dominio con grado di appartenenza maggiore di zero, ossia è l'insieme (questa volta CRISP) di tutti gli elementi per cui non possiamo dire che non appartengano a F (quindi con grado più o meno elevato appartengono a F).

Height of a fuzzy set, F

$$\text{height}(F) = \max_{x \in X} \{f_F(x)\}$$

82
3

Maria Luisa Sapino (BDMM 2010)

L'altezza di un insieme F è il massimo valore di appartenenza di un elemento di F.

Height of a fuzzy set, F

$$\text{height}(F) = \max_{x \in X} \{f_F(x)\}$$

82
4

Maria Luisa Sapino (BDMM 2010)

Un insieme fuzzy normale è un insieme fuzzy con altezza 1.

Height of a fuzzy set, F

$$\text{height}(F) = \max_{x \in X} \{f_F(x)\}$$

Maria Luisa Sapino (BDMM 2010)

82
5

Se invece l'altezza di un insieme fuzzy F è minore di uno, allora F è subnormale.

Normalized fuzzy set, F^*

$$F^* = F / \text{height}(F)$$

Maria Luisa Sapino (BDMM 2010)

82
6

Un insieme fuzzy normalizzato F^* normalizzato è un insieme costruito a partire da un insieme fuzzy F in cui ogni valore di appartenenza è diviso con $\text{height}(F)$ così F^* diventa un insieme normale.

Normalized fuzzy set, F^*

$$F^* = F / \text{height}(F)$$

Maria Luisa Sapino (BDMM 2010)

82
7

- il SUPPORTO dell'insieme normalizzato non cambia perché i valori che erano > 0 rimangono > 0 , e quelli che erano $= 0$ rimangono $= 0$

Cardinality of F

$$card(F) = \sum_{x \in X} f_F(x)$$

$$card(F) = \int_{x \in X} f_F(x) dx$$

82
8

Maria Luisa Sapino (BDMM 2010)

- cardinalità di un insieme fuzzy: somma dei gradi di appartenenza di tutti i suoi elementi, se il dominio è discreto (se è continuo si calcola l'integrale, es. temperatura varia nel continuo)

Cardinality of F

$$card(F) = \sum_{x \in X} f_F(x)$$

$$card(F) = \int_{x \in X} f_F(x) dx$$

82
9

Maria Luisa Sapino (BDMM 2010)

Probabilistic databases:
cardinality is 1

- insieme fuzzy tagliato al grado α : insieme fuzzy ottenuto considerando solo tutti gli elementi il cui grado di appartenenza è $> \alpha$ ---> lo ottengo imponendo una soglia minima di appartenenza, quindi è un'operazione di filtraggio

α -cut a fuzzy set, F

$$F^{>\alpha} = \{x \in X \mid f_F(x) > \alpha\}$$

83
0

Maria Luisa Sapino (BDMM 2010)

Strong α -cut a fuzzy set, F

$$F^{\geq\alpha} = \{x \in X \mid f_F(x) \geq \alpha\}$$

83
1

Maria Luisa Sapino (BDMM 2010)

- se è taglio forte ammetto anche il grado di appartenenza /alpha INCLUSO
-
-
-
-
-

Kernel of F

$$\text{kernel}(F) = F^{\geq 1}$$

83
2

Maria Luisa Sapino (BDMM 2010)

- KERNEL di un insieme fuzzy: insieme degli elementi che appartengono con certezza (quindi taglio forte a livello 1)
-
-
-
-
-

Set operations

$$A \subseteq B \Leftrightarrow \forall x \quad f_A(x) \leq f_B(x)$$

83
3

Maria Luisa Sapino (BDMM 2010)

operazioni insiemistiche:

- A è SOTTOSTIMA di B sse per ogni x il grado di appartenenza di x ad A è minore o uguale del grado di appartenenza di x all'insieme B ---> ciascun elemento appartiene a B almeno quanto appartiene ad A
-
-
-
-

es. A è un disgiunto di B

A = individuo giovane

B = individuo giovane o atletico

- se un individuo appartiene ad A con grado n, allora appartiene a B con grado ALMENO n ---> calcolare il grado di appartenenza ad un disgiunto ci dà una sottostima del grado di appartenenza alla disunione!

Set operations

$$A \subseteq B \Leftrightarrow \forall x \ f_A(x) \leq f_B(x)$$

A underestimates B
Maria Luisa Sapino (BDMM 2010)

83
4

Set operations

$$A \subseteq B \Leftrightarrow A^{>\alpha} \subseteq B^{>\alpha}$$

$$A \subseteq B \Leftrightarrow A^{\geq\alpha} \subseteq B^{\geq\alpha}$$

Maria Luisa Sapino (BDMM 2010)

83
5

Set operations

$$A \subseteq B \Leftrightarrow \text{supp}(A) \subseteq \text{supp}(B)$$

$$A \subseteq B \Leftrightarrow \text{height}(A) \leq \text{height}(B)$$

Maria Luisa Sapino (BDMM 2010)

83
6

vale anche relazione di sottostima tra gli insieme di supporto: al massimo è possibile che elementi in A abbiano grado 0 e in B abbiano grado > 0, ma non viceversa!

se A è sottostima di B, allora l'altezza di A è minore o uguale dell'altezza di B: infatti il valore massimo di A può crescere o rimanere uguale in B

Example query

$Q(X) \leftarrow [s.like] man, X.semantic_property \wedge [image.match](X.image_property, "a.gif").$

83
7

Maria Luisa Sapino (BDMM 2010)

ora possiamo definire gli operatori sugli insiemi fuzzy che utilizziamo nella valutazione delle query: usiamo insiemi fuzzy per quantificare il grado di soddisfacimento di prediciati imperfetti

Example query

$Q(X) \leftarrow [s.like] man, X.semantic_property \wedge [image.match](X.image_property, "a.gif").$

0.84 0.68

83
8

Maria Luisa Sapino (BDMM 2010)

es. supponiamo di avere associato due gradi di soddisfacimento ai prediciati

Example query

$Q(X) \leftarrow [s.like] man, X.semantic_property \wedge [image.match](X.image_property, "a.gif").$

0.76 0.84 0.68
Fuzzy logical operator

83
9

Maria Luisa Sapino (BDMM 2010)

dobbiamo definire come funziona l'operatore di congiunzione fuzzy!

Example query

The diagram illustrates a query expression with annotations:

$$Q(X) \leftarrow \text{man}, X.\text{semantic_property} \text{ ? } \text{image_match}(X.\text{image_property}, "a.gif").$$

Annotations:

- A red bracket labeled "Fuzzy (imperfect)" covers the predicate $X.\text{semantic_property}$.
- A red bracket labeled "Fuzzy logical operator" covers the entire expression $\text{? } \text{image_match}(X.\text{image_property}, "a.gif")$.
- Below the expression, values 0.76, 0.84, and 0.68 are shown under the respective components.

Below the diagram:

$$Q(Y_1, \dots, Y_n) \leftarrow \Theta(p_1(Y_1, \dots, Y_n), \dots, p_m(Y_1, \dots, Y_n)),$$

- Fuzzy and crisp predicates
- Fuzzy logical expression and a merge function
- Results is a ranked list (with the associated fuzzy values!)

Maria Luisa Sapino (BDMM 2010)

How to process a fuzzy query?

The graph shows a membership function for the predicate "hot". The vertical axis is labeled "hot" with values 0.0, 0.7, and 1.0. The horizontal axis is labeled "heat" with values from -5 to 45. A green shaded area represents the membership degree, starting at 0.0 for heat < 0, rising to 0.7 at heat ≈ 5, and reaching 1.0 at heat ≈ 20. A red step function indicates the crispified result, which is 1.0 for heat ≥ 5 and 0.0 for heat < 5.

Maria Luisa Sapino (BDMM 2010)

How to process a fuzzy query?

The graph is identical to the one above, showing the membership function for "hot" and the resulting crispified result. Below the graph, a note states:

Good for processing in traditional databases
Not good for multimedia applications

Maria Luisa Sapino (BDMM 2010)

ci sono diversi modi di calcolare la congiunzione di valori fuzzy, ad es. fare la media dei valori

dato un insieme di predicati p_1, \dots, p_m e dato un operatore logico, dobbiamo associare un grado di plausibilità alla composizione dei predicati parziali tramite l'operatore logico

- dobbiamo consentire la completezza sia di predication fuzzy che crisp
- dobbiamo definire il concetto di espressioni logiche fuzzy e funzioni di merge fuzzy
- dobbiamo vedere come sfruttare le liste ordinate dei risultati parziali per restituire in modo efficiente risultati globali ordinati

1) primo approccio: rendo il predicato CRISP \rightarrow sopra una certa soglia diventa 1, sotto la soglia diventa 0

non ha molto senso tornare alla definizione binaria...

nei dbm i dati sono intrinsecamente fuzzy, quindi è consigliabile non passare a rappresentazione crisp!

How to process a fuzzy query?

$$Q(Y_1, \dots, Y_n) \leftarrow \Theta(p_1(Y_1, \dots, Y_n), \dots, p_m(Y_1, \dots, Y_n)),$$

- Second approach...use suitable fuzzy logic!!!
 - Merge (or scoring) functions....

$$Q(X) \leftarrow [s.likes] man, X.semantic.property) \wedge [image.matched](X.image.property, "a.gif").$$

0.76

0.84

0.68

$$0.76 = \mu_{\wedge}(0.84, 0.68)$$

84
3

Maria Luisa Sapino (BDMM 2010)

2) definiamo opportune funzioni di MERGE che combinano i risultati di predici fuzzy parziali

/mu = Merge

/mu AND = funzione di merge che combina due valori fuzzy ---> in questo caso ne fa la media

Example merge functions...

Min semantics	
$\mu_{P_i \wedge P_j}(x) = \min\{\mu_i(x), \mu_j(x)\}$	
$\mu_{P_i \vee P_j}(x) = \max\{\mu_i(x), \mu_j(x)\}$	
$\mu_{\neg P_i}(x) = 1 - \mu_i(x)$	

84
4

Maria Luisa Sapino (BDMM 2010)

esempi di funzioni di combinazione (definiscono la semantica)

1) SEMANTICA DEL MINIMO:

- $p_1 \text{ AND } p_2^* = \min\{p_1^*, p_2^*\}$
- $p_1 \text{ OR } p_2^* = \max\{p_1^*, p_2^*\}$
- $\text{NOT } p^* = 1 - p^*$

Example merge functions...

Min semantics	
$\mu_{P_i \wedge P_j}(x) = \min\{\mu_i(x), \mu_j(x)\}$	$\mu_{P_i \wedge P_j}(x) = \frac{\mu_i(x) \times \mu_j(x)}{\max(\mu_i(x), \mu_j(x), \alpha)}$ $\alpha \in [0, 1]$
$\mu_{P_i \vee P_j}(x) = \max\{\mu_i(x), \mu_j(x)\}$	$\mu_{P_i \vee P_j}(x) = \frac{\mu_i(x) + \mu_j(x) - \mu_i(x) \times \mu_j(x) - \min(\mu_i(x), \mu_j(x), 1-\alpha)}{\max(1-\mu_i(x), 1-\mu_j(x), \alpha)}$
$\mu_{\neg P_i}(x) = 1 - \mu_i(x)$	$\mu_{\neg P_i}(x) = 1 - \mu_i(x)$

84
5

Maria Luisa Sapino (BDMM 2010)

2) SEMANTICA DEL PRODOTTO:

- $p_1 \text{ AND } p_2^* = p_1^* \times p_2^* / \max\{p_1^*, p_2^*, \alpha\}$
(necessario perché se entrambi sono 0 ho denominatore = 0)

- non vediamo dettagli dell'OR
- NOT rimane uguale al caso del minimo

Example merge functions...

Min semantics	Product semantics
$\mu_{P_i \wedge P_j}(x) = \min\{\mu_i(x), \mu_j(x)\}$	$\mu_{P_i \wedge P_j}(x) = \frac{\mu_i(x) \times \mu_j(x)}{\max\{\mu_i(x), \mu_j(x)\}}, \alpha \in [0, 1]$
$\mu_{P_i \vee P_j}(x) = \max\{\mu_i(x), \mu_j(x)\}$	$\mu_{P_i \vee P_j}(x) = \frac{\mu_i(x) + \mu_j(x) - \mu_i(x) \times \mu_j(x)}{\max\{\mu_i(x), \mu_j(x)\} - \min\{\mu_i(x), \mu_j(x)\}, \alpha}$
$\mu_{\neg P_i}(x) = 1 - \mu_i(x)$	$\mu_{\neg P_i}(x) = 1 - \mu_i(x)$

Arithmetic average (N-ary)		
$\mu_{P_1 \wedge \dots \wedge P_n}(x)$	$\mu_{\neg P_i}(x)$	$\mu_{P_1 \vee \dots \vee P_n}(x)$
$\frac{\mu_1(x) + \dots + \mu_n(x)}{ \{P_1, \dots, P_n\} }$	$1 - \mu_i(x)$	$\frac{ \{P_1, \dots, P_n\} - \mu_i(x) + \dots + \mu_j(x)}{ \{P_1, \dots, P_n\} }$
(mostly used in information retrieval)		

84
6

Maria Luisa Sapino (BDMM 2010)

3) SEMANTICA DELLA MEDIA

- $p_1 \text{ AND } \dots \text{ AND } p_n^* = (p_1^* + p_n^*) / n$
- $p_1 \text{ OR } p_2^* = (n - P_1^* + \dots + P_n^*) / n$
- $\text{NOT } p^* = 1 - p^*$

Il caso dell'or sembra non avere molto senso così com'è scritto!

Triangular norms (and co-norms)

- How to emulate the properties of a crisp predicate

	T-norm binary function N (for \wedge)	T-conorm binary function C (for \vee)
Boundary conditions	$N(0, 0) = 0, N(x, 1) = N(1, x) = x$	$C(1, 1) = 1, C(x, 0) = C(0, x) = x$

84
7

Maria Luisa Sapino (BDMM 2010)

dobbiamo poter gestire anche prediciati crisp (esatti), quindi
dobbiamo preoccuparci dei casi limiti min e max, cioè 0 e 1 che sono gli unici valori assunti dai prediciati crisp

---> funzioni f che soddisfano i requisiti di norma o co-norma:

FUNZIONE NORMA f (casi limite per AND):

$$* f(0, 0) = 0$$

$$* f(1, x) = f(x, 1) = x$$

il MINIMO e il PRODOTTO soddisfano queste proprietà

FUNZIONE CO-NORMA f (casi limite per OR)

$$* f(1, 1) = 1$$

$$* f(0, x) = f(x, 0) = x$$

la funzione MASSIMO soddisfa il requisito

---> la semantica MINIMO soddisfa tutti i requisiti

deve valere la commutatività (vale per min-max)

Triangular norms (and co-norms)

- How to emulate the properties of a crisp predicate

	T-norm binary function N (for \wedge)	T-conorm binary function C (for \vee)
Boundary conditions	$N(0, 0) = 0, N(x, 1) = N(1, x) = x$	$C(1, 1) = 1, C(x, 0) = C(0, x) = x$
Commutativity	$N(x, y) = N(y, x)$	$C(x, y) = C(y, x)$

84
8

Maria Luisa Sapino (BDMM 2010)

Triangular norms (and co-norms)

- How to emulate the properties of a crisp predicate

	T-norm binary function N (for \wedge)	T-conorm binary function C (for \vee)
Boundary conditions	$N(0,0) = 0, N(x,1) = N(1,x) = x$	$C(1,1) = 1, C(x,0) = C(0,x) = x$
Commutativity	$N(x,y) = N(y,x)$	$C(x,y) = C(y,x)$
Monotonicity	$x \leq x', y \leq y' \rightarrow N(x,y) \leq N(x',y')$	$x \leq x', y \leq y' \rightarrow C(x,y) \leq C(x',y')$

84
9

Maria Luisa Sapino (BDMM 2010)

- deve valere monotonicità: la combinazione di sottostime di x' e y' deve essere una sottostima della combinazione di x' e y'

Triangular norms (and co-norms)

- How to emulate the properties of a crisp predicate

	T-norm binary function N (for \wedge)	T-conorm binary function C (for \vee)
Boundary conditions	$N(0,0) = 0, N(x,1) = N(1,x) = x$	$C(1,1) = 1, C(x,0) = C(0,x) = x$
Commutativity	$N(x,y) = N(y,x)$	$C(x,y) = C(y,x)$
Monotonicity	$x \leq x', y \leq y' \rightarrow N(x,y) \leq N(x',y')$	$x \leq x', y \leq y' \rightarrow C(x,y) \leq C(x',y')$
Associativity	$N(x,N(y,z)) \leq N(N(x,y),z)$	$C(x,C(y,z)) \leq C(C(x,y),z)$

85
0

Maria Luisa Sapino (BDMM 2010)

- deve valere associatività, quindi se devo combinare 3 elementi non importa l'ordine (vale per max-min)

Triangular norms (and co-norms)

- How to emulate the properties of a crisp predicate

	T-norm binary function N (for \wedge)	T-conorm binary function C (for \vee)
Boundary conditions	$N(0,0) = 0, N(x,1) = N(1,x) = x$	$C(1,1) = 1, C(x,0) = C(0,x) = x$
Commutativity	$N(x,y) = N(y,x)$	$C(x,y) = C(y,x)$
Monotonicity	$x \leq x', y \leq y' \rightarrow N(x,y) \leq N(x',y')$	$x \leq x', y \leq y' \rightarrow C(x,y) \leq C(x',y')$
Associativity	$N(x,N(y,z)) \leq N(N(x,y),z)$	$C(x,C(y,z)) \leq C(C(x,y),z)$

85
1

Bellman and Giertz: "The unique aggregation functions for evaluating AND and OR that preserve logical equivalence of queries involving only conjunction and disjunction and that are monotonic in their arguments are min and max."

Maria Luisa Sapino (BDMM 2010)

MIN-MAX sono diffusi (nonostante i loro limiti) perché per formule che contengono solo AND e OR sono gli unici operatori che preservano l'equivalenza logica, ovvero se applicati a formule logicamente equivalenti danno lo stesso risultato --> se garantisco l'equivalenza logica, allora posso usare i risultati parziali per ottimizzare

(per la MEDIA non funziona!)

Triangular norms (and co-norms)

- Emulating the properties of a crisp predicate may not be good for multimedia applications!!!
- Boundary conditions prevent partial matches

$Q(X) \leftarrow \text{s.likes}(man, X.\text{semantic_property}) \wedge \text{image.matched}(X.\text{image_property}, "a.gif")$

0.00	0.99	0.00
------	------	------

$0.00 = \mu_{\text{L,min}}(0.99, 0.00)$

Maria Luisa Sapino (BDMM 2010)

85
2

Triangular norms (and co-norms)

- Emulating the properties of a crisp predicate may not be good for multimedia applications!!!
- Monotone condition is weak!!

$Q(X) \leftarrow \text{s.likes}(man, X.\text{semantic_property}) \wedge \text{image.matched}(X.\text{image_property}, "a.gif")$

0.70	0.71	0.70
0.70	0.99	0.70

$0.70 = \mu_{\text{L,min}}(0.71, 0.70)$

$0.70 = \mu_{\text{L,min}}(0.99, 0.70)$

Maria Luisa Sapino (BDMM 2010)

85
3

Triangular norms (and co-norms)

- Emulating the properties of a crisp predicate may not be good for multimedia applications!!!
- N-ary semantics may be enough !!!
 - consider all relevant features at the same time, instead of in pairs!

Arithmetic average (N-ary)

$\mu_{P_1 \wedge \dots \wedge P_n}(x)$	$\mu_{\neg P_i}(x)$	$\mu_{P_1 \vee \dots \vee P_n}(x)$
$\frac{\mu_1(x) + \dots + \mu_n(x)}{ (P_1, \dots, P_n) }$	$1 - \mu_i(x)$	$\frac{ \{(P_1, \dots, P_n)\} - \mu_1(x) - \dots - \mu_n(x)}{ (P_1, \dots, P_n) }$

Geometric average (N-ary)

$\mu_{P_1 \wedge \dots \wedge P_n}(x)$	$\mu_{\neg P_i}(x)$	$\mu_{P_1 \vee \dots \vee P_n}(x)$
$(\mu_1(x) \times \dots \times \mu_n(x))^{\frac{1}{n}}$	$1 - \mu_i(x)$	$1 - ((1 - \mu_1(x)) \times \dots \times (1 - \mu_n(x)))^{\frac{1}{n}}$

IM 2010

85
4

nelle dbm il fatto di emulare nelle condizioni limite i risultati di prediciati crisp non è necessariamente un requisito desiderabile: può succedere che le condizioni limite neutralizzino il valore degli altri risultati parziali!

nell'es. usando la semantica del minimo ottengo 0 anche se il primo requisito è soddisfatto quasi al massimo, quindi con questa semantica non potrei distinguere da un altro risultato in cui entrambi i parziali sono 0 (e che quindi non soddisfa per niente la query)!

---> MAX-MIN va bene per equivalenza logica ma
-) nasconde i risultati parziali

-) soddisfa la condizione di monotonicità ma non consente di discriminare abbastanza, nell'es. il primo risultato soddisfa meno del secondo ma il risultato globale è lo stesso!

---> se invece uso la MEDIA perdo l'equivalenza logica ma ho maggiore capacità di discriminare

posso considerare semantica n-aria per combinare più prediciati con lo stesso operatore

- MEDIA ARITMETICA
- MEDIA GEOMETRICA

Visualisation of minimum

(c) **Composition - Minimum**

Maria Luisa Sapino (BDMM 2010)

85
8

visualizzazione dell'AND nel caso della semantica MINIMO

Visualisation of minimum

(c) **Composition - Minimum**

Maria Luisa Sapino (BDMM 2010)

85
9

-) è monotono ma ad ogni punto (x, y) , per una delle due coordinate (es. x) corrisponde un'intera fascia di valori (es. da x a 1) per cui la combinazione dei valori è uguale, ovvero non influisce sul risultato finale

-) non è sufficiente a garantire il match parziale: se anche solo uno dei due ha valore 0, allora l'altro non contribuisce per niente al risultato globale

Visualisation of arithmetic average

(c) **Composition - Arithmetic Average**

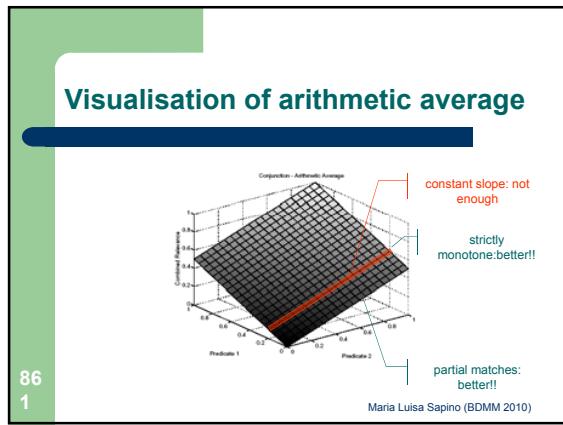
Maria Luisa Sapino (BDMM 2010)

86
0

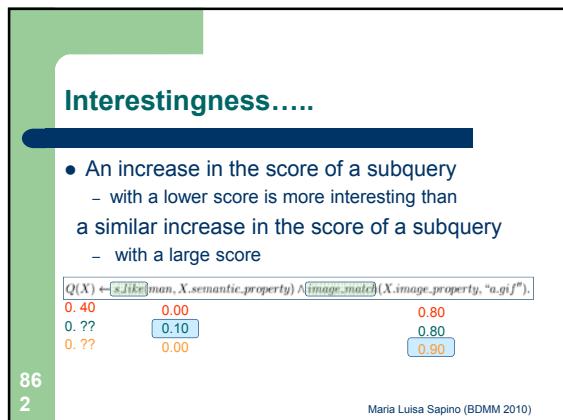
visualizzazione dell'AND nel caso della MEDIA ARITMETICA

+) vale 0 solo se entrambi hanno valore 0 ed è commutativa

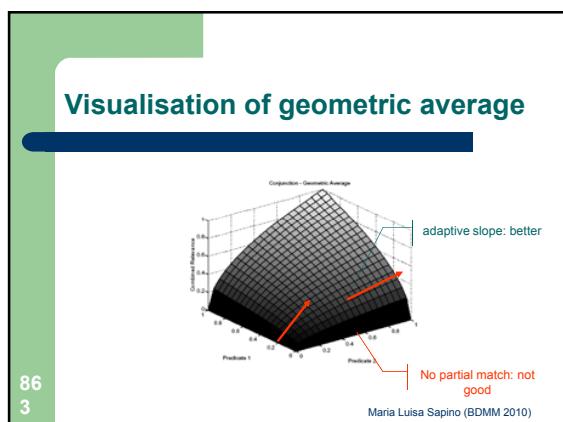
+) meglio anche per il match parziale: se almeno uno dei due è diverso da 0, allora contribuisce al risultato globale



-) incremento di valore di una delle due coordinate ha impatto sulla media allo stesso modo a prescindere dal livello in cui è posizionato, es. incremento da 0 a 0.2 e incremento da 0.2 a 0.4 pesano allo stesso modo, ma in certi casi si vorrebbe premiare un incremento dal basso



in questo caso potrei voler preferire il secondo risultato, perché un risultato parziale è passato da avere peso 0 ad averne 0.1



visualizzazione dell'AND nel caso della MEDIA GEOMETRICA

- +) crescita più ripida per valori più bassi, quindi privilegia crescita di valori bassi
 -) non c'è match parziale (come MINIMO)
-
-
-
-

....parametric geometric average

$$\mu_{(P_1 \wedge \dots \wedge P_n)}(t, r_{true}, \beta) = \frac{((\prod_{k=1}^n \mu_k(t) \geq r_{true}) \times (\prod_{k=1}^n \mu_k(t) < r_{true}) \beta)^{1/n} - \beta}{1 - \beta}$$

Truth cutoff Falsehood value

86
4

Maria Luisa Sapino (BDMM 2010)

come migliorare questo aspetto negativo della media geometrica?

MEDIA GEOMETRICA PARAMETRIZZATA

media geometrica in cui definiamo una "soglia di verità" al di sotto della quale i predicati sono considerati falsi:

se un predicato raggiunge la soglia allora interviene nel prodotto esattamente con il suo peso, altrimenti interviene con peso β --> quindi β diventa il valore minimo di verità (anche se magari in origine era un valore inferiore a β)

....parametric geometric average

$$\mu_{(P_1 \wedge \dots \wedge P_n)}(t, r_{true}, \beta) = \frac{((\prod_{k=1}^n \mu_k(t) \geq r_{true}) \times (\prod_{k=1}^n \mu_k(t) < r_{true}) \beta)^{1/n} - \beta}{1 - \beta}$$

r_{true}=0.4; β=0.4 r_{true}=0.4; β=0.2 r_{true}=0.4; β=0.0

86
5

Maria Luisa Sapino (BDMM 2010)

visualizzazione dell'AND nel caso della MEDIA GEOMETRICA PARAMETRIZZATA

How to put weights?????

- How do I state that image properties are more important than semantic properties??
- What do we mean?:
 - A change in the value of image property should have a larger impact than a similar change in the value of the semantic property.

$Q(X) \leftarrow \text{s.likes}(man, X.\text{semantic_property}) \wedge \text{image.match}(X.\text{image_property}, "a.gif")$		
0.60	0.60	0.60
0.65	0.70	0.60
0.68	0.60	0.70

86
6

Maria Luisa Sapino (BDMM 2010)

come indichiamo che alcune proprietà pesano più di altre?

How to put weights?????

$Q(X) \leftarrow [s.like]man, X.semantic_property] \wedge [image_match](X.image_property, "a.gif").$

- How do I state that image properties are more important than semantic properties??

86
7

Maria Luisa Sapino (BDMM 2010)

Fagin's proposal

- Desiderata

- If all weights are equal the result should be equal to no weight case

86
8

Maria Luisa Sapino (BDMM 2010)

Fagin ha formulato un insieme di requisiti (DESIDERATA) che un buon sistema di pesatura deve soddisfare:

1) se i pesi sono tutti uguali, allora il risultato della valutazione deve essere lo stesso che avrei in assenza di pesi

Fagin's proposal

- Desiderata

- If all weights are equal the result should be equal to no weight case
- If one of the weights is zero, the subquery can be dropped without effecting the rest

86
9

Maria Luisa Sapino (BDMM 2010)

2) se uno dei pesi è 0, allora la sottoquery che ha quel peso può essere eliminata senza avere impatto sul risultato finale --> corrisponde a introdurre un criterio di valutazione che non viene considerato nel risultato finale

Fagin's proposal

- Desirata
 - If all weights are equal the result should be equal to no weight case
 - If one of the weights is zero, the subquery can be dropped without effecting the rest
 - ..the result should be a continuous function of the weights

87
0

Maria Luisa Sapino (BDMM 2010)

3) il risultato deve essere una funzione continua dei pesi

Fagin's proposal

- Let $\theta_1 + \theta_2 + \dots + \theta_m = 1$
 $\theta_1, \theta_2, \dots, \theta_m \geq 0$
 $\theta_1 \geq \theta_2 \geq \dots \geq \theta_m$

87
1

- i pesi sommano a 1 (basta normalizzare)
- i pesi devono essere tutti $> 0 = a 0$
- i pesi sono disposti in ordine decrescente: non è un'ipotesi rigida --> si può fare perché stiamo considerando la congiunzione che è commutativa, quindi posso sempre riformulare la query in modo da portare davanti la sottoquery che pesa di più

Fagin's proposal

- Let $\theta_1 + \theta_2 + \dots + \theta_m = 1$
 $\theta_1, \theta_2, \dots, \theta_m \geq 0$
 $\theta_1 \geq \theta_2 \geq \dots \geq \theta_m$
- then $f_{(\theta_1, \theta_2, \dots, \theta_m)}(x_1, x_2, \dots, x_m) = (\theta_1 - \theta_2)f(x_1) + 2(\theta_2 - \theta_3)f(x_1, x_2) + 3(\theta_3 - \theta_4)f(x_1, x_2, x_3) + \dots + (m-1)(\theta_{(m-1)} - \theta_m)f(x_1, x_2, \dots, x_{(m-1)}) + m\theta_m f(x_1, x_2, x_3, \dots, x_m)$

87
2

funzione di combinazione con pesi \Theta_1..m applicata a m predici:

- n righe
- per ogni riga:

#riga * (\Theta_#riga - \Theta_{#riga+1}) * f(x_1, ..., f_#riga)

- allora la funzione di composizione equivale alla sommatoria per tutti i valori da 1 a n delle righe calcolate sopra

--> QUESTO SCHEMA SODDISFA I REQUISITI DI FAGIN!

Fagin's proposal

- Let $\theta_1 + \theta_2 + \dots + \theta_m = 1$
 $\theta_1, \theta_2, \dots, \theta_m \geq 0$
 $\theta_1 \geq \theta_2 \geq \dots \geq \theta_m$
 - then $f_{(\theta_1, \theta_2, \dots, \theta_m)}(x_1, x_2, \dots, x_m) = (\theta_1 - \theta_2)f(x_1) + 2(\theta_2 - \theta_3)f(x_1, x_2) + 3(\theta_3 - \theta_4)f(x_1, x_2, x_3) + \dots + (m-1)(\theta_{(m-1)} - \theta_m)f(x_1, x_2, x_3, \dots, x_{(m-1)}) + m\theta_m f(x_1, x_2, x_3, \dots, x_m)$

If f is continuous, then the weighted function is also continuous

87
3

- è continua se f è continua

Fagin's proposal

- Let $\theta_1 + \theta_2 + \dots + \theta_m = 1$
 $\theta_1, \theta_2, \dots, \theta_m \geq 0$
 $\theta_1 \geq \theta_2 \geq \dots \geq \theta_m$
 - then $f_{(\theta_1, \theta_2, \dots, \theta_m)}(x_1, x_2, \dots, x_m) = (\theta_1 - \theta_2)f(x_1) + 2(\theta_2 - \theta_3)f(x_1, x_2) + 3(\theta_3 - \theta_4)f(x_1, x_2, x_3) + \dots + (m-1)\theta_{(m-1)}f(x_1, x_2, x_3, \dots, x_{(m-1)})$

If lowest weight is 0, then the corresponding sub query can be omitted

87
4

- se uno dei pesi è 0 tanto vale ignorarli: visto che sono ordinati è sicuramente l'ultimo valore --> se $m=0$ l'ultimo termine si annulla, quindi ottengo esattamente la definizione della stessa funzione senza l'elemento m

Fagin's proposal

- Let $\theta_1 + \theta_2 + \dots + \theta_m = 1$
 $\theta_1, \theta_2, \dots, \theta_m \geq 0$
 $\theta_1 \geq \theta_2 \geq \dots \geq \theta_m$
 - then $f_{\left(\frac{1}{m}, \frac{1}{m}, \dots, \frac{1}{m}\right)}(x_1, x_2, \dots, x_m) = \left(\frac{1}{m} - \frac{1}{m}\right)f(x_1) + 2\left(\frac{1}{m} - \frac{1}{m}\right)f(x_1, x_2) + \dots +$
If all weights are equal... $3\left(\frac{1}{m} - \frac{1}{m}\right)f(x_1, x_2, x_3) + \dots +$
 $(m-1)\left(\frac{1}{m} - \frac{1}{m}\right)f(x_1, x_2, x_3, \dots, x_{(m-1)}) +$
 $m\frac{1}{m}f(x_1, x_2, x_3, \dots, x_m)$

87
5

- se tutti i pesi sono uguali, allora tutti i termini si azzerano tranne l'ultimo, che corrisponde alla funzione senza pesi

Fagin's proposal

- Let $\theta_1 + \theta_2 + \dots + \theta_m = 1$
 $\theta_1, \theta_2, \dots, \theta_m \geq 0$
 $\theta_1 \geq \theta_2 \geq \dots \geq \theta_m$
- then $f_{\left(\frac{1}{m}, \frac{1}{m}, \dots, \frac{1}{m}\right)}(x_1, x_2, \dots, x_m) = f(x_1, x_2, x_3, \dots, x_m)$

If all weights are equal then the result is equal to the no-weighted case

87
6

Maria Luisa Sapino (BDMM 2010)

Example (arithmetic average)

$$\text{score}(a \wedge b) = \frac{\text{score}(a) + \text{score}(b)}{2}$$

87
7

Maria Luisa Sapino (BDMM 2010)

es. di applicazione della formula di Fagin alla media aritmetica

Example (arithmetic average)

$$\text{score}(a \wedge b) = \frac{\text{score}(a) + \text{score}(b)}{2}$$

$$\text{score}(a \wedge b) = (\theta_a - \theta_b)\text{score}(a) + 2\theta_b \frac{\text{score}(a) + \text{score}(b)}{2}$$

87
8

Maria Luisa Sapino (BDMM 2010)

supponiamo di voler pesare diversamente le due componenti: allora applichiamo la formula di Fagin usando $f = \text{media aritmetica}$

Example (arithmetic average)

$$\text{score}(a \wedge b) = \frac{\text{score}(a) + \text{score}(b)}{2}$$

$$\text{score}(a \wedge b) = (\theta_a - \theta_b)\text{score}(a) + 2\theta_b \frac{\text{score}(a) + \text{score}(b)}{2}$$

$$\text{score}(a \wedge b) = \theta_a \text{score}(a) + \theta_b \text{score}(b)$$

87
9

Maria Luisa Sapino (BDMM 2010)

si può applicare anche al prodotto

Example (product)

$$\text{score}(a \wedge b) = \text{score}(a) \times \text{score}(b)$$

$$\text{score}(a \wedge b) = (\theta_a - \theta_b)\text{score}(a) + 2\theta_b \text{score}(a) \times \text{score}(b)$$

88
0

Maria Luisa Sapino (BDMM 2010)

PROBLEMA:

desiderata di Fagin non tengono conto delle derivate parziali che esprimono la velocità con cui cresce la funzione combinata rispetto ai punti del dominio in cui avviene il cambiamento --> vogliamo pesare di più i cambiamenti a livello basso

Are Fagin's desiderata enough?

- It does not compare partial derivatives!!!

88
1

Maria Luisa Sapino (BDMM 2010)

Are Fagin's desiderata enough?

- It does not compare partial derivatives!!!

The figure shows a 3D surface plot representing a function of two variables, $Predicate 1$ and $Predicate 2$, both ranging from 0 to 1. The vertical axis represents the function value, also ranging from 0 to 1. The surface is a smooth, curved plane that starts at (0,0) with a value of 0 and rises to (1,1) with a value of 1. Two vectors originate from the point (0,0): one red vector points along the surface towards the point (1,0), and another blue vector points vertically upwards, representing the adaptive slope. A legend on the right side of the plot area identifies these vectors as "adaptive slope:" and "adaptive importance".

88

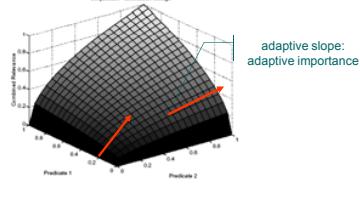
2

Maria Luisa Sapino (BDMM 2010)

88
2

Are Fagin's desiderata enough?

- It does not compare partial derivatives!!!



Maria Luisa Sapino (BDMM 2010)

Are Fagin's desiderata enough?

- It does not compare partial derivatives!!!

88
3

Are Fagin's desiderata enough?

- It does not compare partial derivatives!!!

$$\boxed{E = 1 - \frac{\frac{1}{b^2} + \frac{b^2}{R}}{\frac{1}{R} + \frac{1}{P}}}$$

Maria Luisa Sapino (BDMM 2010)

misuriamo l'EFFICACIA della variazione (non vediamo come)

Precision/Recall è costante e ha valore b

se e solo se

la derivata della funzione parziale rispetto a R è uguale alla derivata parziale rispetto a P

---> b esprime il rapporto di importanza tra P e R

misuriamo il contributo di x e y alla crescita della funzione di combinazione:

x contribuisce di più di y

se e solo se

per ciascuna istanziazione della funzione (= a è grado di soddisfacimento di x e b è grado di soddisfacimento di y), la derivata parziale rispetto a x è > della derivata parziale rispetto a y sugli stessi punti

Are Fagin's desiderata enough?

88
4

Are Fagin's desiderata enough?

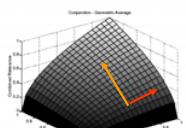
- It does not compare partial derivatives!!!
 - Importance: Given a function $f(x,y)$, x has a higher contribution than y iff

$$\forall a, b \quad \frac{\partial f}{\partial x} \Big|_{(a,b)} > \frac{\partial f}{\partial y} \Big|_{(a,b)}$$

Maria Luisa Sapino (BDMM 2010)

Are Fagin's desiderata enough?

- It does not compare partial derivatives!!!
- Importance: Given a function $f(x,y)$, x has a higher contribution than y iff

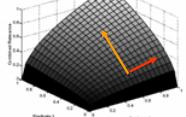
$$\forall a,b \frac{\partial f}{\partial x}|_{(a,b)} > \frac{\partial f}{\partial y}|_{(a,b)}$$


Maria Luisa Sapino (BDMM 2010)

88
5

Are Fagin's desiderata enough?

- It does not compare partial derivatives!!!
- Importance: Given a function $f(x,y)$, x has a higher contribution than y iff

$$\forall a,b \frac{\partial f}{\partial x}|_{(a,b)} > \frac{\partial f}{\partial y}|_{(a,b)}$$


$$\text{relimp}(x,y)|_{(a,b)} = \frac{\frac{\partial f}{\partial x}|_{(a,b)}}{\frac{\partial f}{\partial y}|_{(a,b)}}$$

Maria Luisa Sapino (BDMM 2010)

88
6

Are Fagin's desiderata enough?

- Importance: Given a function $f(x,y)$, x has a higher contribution than y iff $\forall a,b \frac{\partial f}{\partial x}|_{(a,b)} > \frac{\partial f}{\partial y}|_{(a,b)}$
- Example:

$$\text{score}(x \wedge y) = \theta_x \text{score}(x) + \theta_y \text{score}(y)$$

$$\frac{\partial \text{score}(x \wedge y)}{\partial \text{score}(x)} \Big|_{(a,b)} = \theta_x \Big|_{(a,b)} = \theta_x$$

OK!

$$\frac{\partial \text{score}(x \wedge y)}{\partial \text{score}(y)} \Big|_{(a,b)} = \theta_y \Big|_{(a,b)} = \theta_y$$

Maria Luisa Sapino (BDMM 2010)

88
7

relimp = relative importance
--> se x è più importante di y allora relimp > 1,
altrimenti è = 1

in questo caso la derivata parziale mi dà
esattamente il peso relativo che io voglio dare a
quella componente, quindi relimp rispetta la
semantica che volevo

Are Fagin's desiderata enough?

- Importance: Given a function $f(x,y)$, x has a higher contribution than y iff $\forall a,b \frac{\partial f}{\partial x}|_{(a,b)} > \frac{\partial f}{\partial y}|_{(a,b)}$
- Example:

$$score(x \wedge y) = (\theta_x - \theta_y)score(x) + 2\theta_y score(x) \times score(y)$$

$$\frac{\partial score(x \wedge y)}{\partial score(x)} \Big|_{(a,b)} = (\theta_x - \theta_y) + 2\theta_y b$$

NOT OK!

$$\frac{\partial score(x \wedge y)}{\partial score(y)} \Big|_{(a,b)} = 2\theta_y a$$

Maria Luisa Sapino (BDMM 2010)

88
8

===== nel caso del prodotto però non funziona!!! =====

usando come funzione di combinazione il prodotto, il rapporto tra i pesi non è mantenuto dal rapporto tra le derivate parziali, che non è nemmeno costante per cui a seconda dei punti a cui lo applico ho un risultato diverso

Are Fagin's desiderata enough?

- Importance: Given a function $f(x,y)$, x has a higher contribution than y iff $\forall a,b \frac{\partial f}{\partial x}|_{(a,b)} > \frac{\partial f}{\partial y}|_{(a,b)}$
- Example:

$$score(x \wedge y) = (\theta_x - \theta_y)score(x) + 2\theta_y score(x) \times score(y)$$

$$\frac{\partial score(x \wedge y)}{\partial score(x)} \Big|_{(a,b)} = (\theta_x - \theta_y) + 2\theta_y b$$

NOT OK!

$$\frac{\partial score(x \wedge y)}{\partial score(y)} \Big|_{(a,b)} = 2\theta_y a$$

Maria Luisa Sapino (BDMM 2010)

88
9

Are Fagin's desiderata enough?

- Importance: Given a function $f(x,y)$, x has a higher contribution than y iff $\forall a,b \frac{\partial f}{\partial x}|_{(a,b)} > \frac{\partial f}{\partial y}|_{(a,b)}$
- Example:

$$score(x \wedge y) = (\theta_x - \theta_y)score(x) + 2\theta_y score(x) \times score(y)$$

$$\frac{\partial score(x \wedge y)}{\partial score(x)} \Big|_{(a,b)} = (\theta_x - \theta_y) + 2\theta_y b$$

NOT OK!

$$\frac{\partial score(x \wedge y)}{\partial score(y)} \Big|_{(a,b)} = 2\theta_y a$$

Maria Luisa Sapino (BDMM 2010)

89
0

se uso lo stesso grado, l'importanza relativa è > 1
---> quindi la prima pesa di più!

$$\text{relimp}(x,y)|_{\langle a,b \rangle} = \frac{\theta_x b}{\theta_y a}$$

How about?

- Importance: Given a function $f(x,y)$, x has a higher contribution than y iff $\forall a,b \frac{\partial f}{\partial x} \Big|_{(a,b)} > \frac{\partial f}{\partial y} \Big|_{(a,b)}$
 - Example:

$$score(x \wedge y) = score(x)^{\theta_x} \times score(y)^{\theta_y}$$

$$\left. \frac{\partial \text{score}(x \wedge y)}{\partial \text{score}(x)} \right|_{(a,b)} = \theta_x a^{\theta_x - 1} b^{\theta_y}$$

$$\left. \frac{\partial score(x \wedge y)}{\partial score(y)} \right|_{\langle a, b \rangle} = \theta_y a^{\theta_x} b^{\theta_y - 1}$$

Maria Luisa Sapino (BDMM 2010)

$$\text{relimp}(x,y)|_{\langle a,a \rangle} = \frac{\theta_x a}{\theta_y a} = \frac{\theta_x}{\theta_y} > 1$$

How about?

- Importance: Given a function $f(x,y)$, x has a higher contribution than y iff $\forall a,b \frac{\partial f}{\partial x} \Big|_{(a,b)} > \frac{\partial f}{\partial y} \Big|_{(a,b)}$
 - Example:

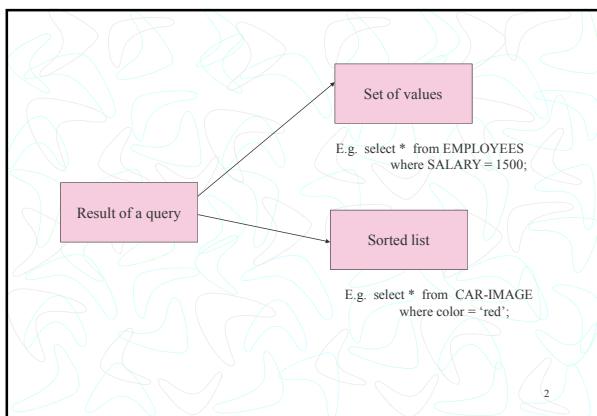
$$score(x \wedge y) = score(x)^{\theta_x} \times score(y)^{\theta_y}$$

$$\left. \frac{\partial score(x \wedge y)}{\partial score(x)} \right|_{\langle a, b \rangle} = \theta_x a^{\theta_x - 1} b^{\theta_y}$$

$$\left. \frac{\partial score(x \wedge y)}{\partial score(y)} \right|_{\langle a, b \rangle} = \theta_y a^{\theta_x} b^{\theta_y - 1}$$

Maria Luisa Sapino (BDMM 2010)





risultato di una query:

* nei DBR è un insieme quindi non è ordinato

* nei DBM è una lista ordinata di valori in base al soddisfacimento della query

Queries as boolean combinations

- How do we combine sets and sorted lists :
- **example:** database containing information about CDs
- query: return the names of all albums whose artist is the *Queen*, and whose cover is mostly *black*.
- (i) traditional database query, asking for the names of all albums whose artist is "Queen"
---> set
- (ii) multimedia query asking for album covers which are "black"
---> sorted list
- (Artist = 'Queen') and (AlbumColor = 'black')

3

query: tipicamente sono combinazioni di predici fuzzy e crisp

Queries as boolean combinations

- What if we replace *and* with *or*?
 - Is the answer a set, a list, or a combination of the two?
- How about if we combine two multimedia queries?
 - (color = ‘black’) and (title contains “kind”)
 - the answer is a sorted list... sorted according to which criterion?

4

Graded sets

- sets of pairs (x, g), where
 - x is an object (such as a tuple)
 - g is a real number in [0,1]
- graded sets as sorted lists, where the objects are sorted by their grades
- graded sets of generalization of both sets and sorted lists

5

usiamo il modello degli insieme fuzzy perché ci consente di rappresentare bene sia insiemi crisp che insiemi fuzzy:
GRADED SET sono insiemi di coppie in cui ogni elemento
che appartiene all’insieme è associato ad un valore
numerico che è il grado con cui l’elemento appartiene
all’insieme

- se ordino gli elementi in base al grado ottengo la lista ordinata che è il risultato del predicato fuzzy

- se uso solo l’elemento quando $g=1$ ho l’insieme crisp

---> ho un unico formalismo che consente di rappresentare entrambe le realtà

Assumptions

- All data in all the subsystems deal with the attributes of a specific set of objects of some fixed types
- Atomic queries are of the form $X = t$, where X is the name of an attribute, and t is a target
- Queries are boolean combination of atomic queries
- For each atomic query, a grade is assigned to each object

6

- abbiamo diversi sottosistemi che gestiscono i vari attributi degli oggetti (es. uno per texture e uno per colori), quindi ogni sottosistema gestisce tutti i dati
- una query atomica è nella forma "proprietà = valore", es.
"color = #000000"
- le query complesse sono combinazioni booleane di query atomiche
- per ogni query atomica si può assegnare un grado di soddisfacimento ad ogni oggetto

Dealing with boolean combinations

- There are a number of **aggregation functions** that assign a grade to a fuzzy conjunction, as a function of the grades assigned to the conjuncts
- Standard rules of fuzzy logic (Zadeh, 1965) use min as the aggregation function:
 - conjunction rule: $\mu_{A \wedge B}(x) = \min\{\mu_A(x), \mu_B(x)\}$
 - disjunction rule : $\mu_{A \vee B}(x) = \max\{\mu_A(x), \mu_B(x)\}$
 - negation rule: $\mu_{\text{not } A}(x) = 1 - \mu_A(x)$

7

qui usiamo la semantica del MINIMO

Standard fuzzy rules: pros

- They are a conservative extension of the standard propositional semantics
- Theorem of Bellman and Giertz:
 - *"The unique aggregation functions for evaluating AND and OR that preserve logical equivalence of queries involving only conjunction and disjunction and that are monotonic in their arguments are min and max."*

8

+) semantica del MINIMO è l'unica che preserva
l'equivalenza logica ed è monotona

Logical equivalence preservation

If Q_1 and Q_2 are logically equivalent queries involving only conjunction and disjunction (not negation), then

$$\mu_{Q_1}(x) = \mu_{Q_2}(x), \quad \text{for every } x$$

Examples:

$$\mu_{A \wedge A}(x) = \mu_A(x)$$

$$\mu_{A \wedge (B \vee C)}(x) = \mu_{(A \wedge B) \vee (A \wedge C)}(x)$$

9

es. valgono tutte le equivalenze logiche:

$$\begin{aligned} * \ A \text{ AND } A &= A \\ * \ A \text{ AND } (B \text{ OR } C) &= (A \text{ AND } B) \text{ OR } (A \text{ AND } C) \end{aligned}$$

questo ci permette di fare ottimizzazioni, ad es. se nel
secondo caso abbiamo già risultati parziali di
A AND B e A AND C

Monotonicity

$$\mu_A(x) \leq \mu_A(x') \text{ and } \mu_B(x) \leq \mu_B(x') \Rightarrow \\ \mu_{A \wedge B}(x) \leq \mu_{A \wedge B}(x')$$

10

* monotonicità:

se
l'elemento x soddisfa la proprietà A non più di quanto la soddisfi x'
e
l'elemento x soddisfa la proprietà B non più di quanto la soddisfi x'
allora
l'elemento x soddisfa la congiunzione di A AND B non più di quanto
la soddisfi x'

es: se un'immagine x è meno rossa e meno gialla di un'altra
immagine x', allora x è complessivamente meno rossa e gialla di x'

Algorithms for query evaluation

- Different subsystems are involved in query answering, and information coming from them is pieced together by the *middleware system* (e.g., Garlic [Fagin])
- (Artist = 'Queen') \wedge (AlbumColor = 'black')
 - Assumption: there are not too many objects that satisfy the first conjunct
 - Intuitive algorithm:
 - Compute S = set of objects that do satisfy (Artist = 'Queen')
 - using random access, obtain grades (e.g. from QBIC) for the objects in S
 - sort objects in S according to the returned grades

11

abbiamo query che è una congiunzione di predicato crisp e
predicato fuzzy

algoritmo intuitivo:

- posso risolvere query crisp che è più facile
- per ciascuno dei risultati del predicato crisp faccio
 - RANDOM ACCESS (=accesso diretto)
 - ordino in base al valore fuzzy degli oggetti

Algorithms for query evaluation

- (AlbumColor = 'black') \wedge (Shape = 'circle')
 - assumption: one subsystem deals with colors, and another one deals with shapes.
 - The grade of any object x under the query is the minimum of the grades of x under the subqueries
 - A1= (AlbumColor = 'black'), and
 - A2= (Shape = 'circle')
- how to get TOP k elements?

12

se però ho solo condizioni atomiche fuzzy non ho a priori garanzie di filtro, quindi non posso usare lo stesso approccio!

- assumiamo di avere sottosistemi che si occupano dei vari attributi
 - usiamo semantica del MINIMO

TOP k naive algorithm

- Have the subsystem dealing with **color** to output explicitly the graded set containing all pairs $(x, \mu_{A1}(x))$, for each x
- Have the subsystem dealing with **shape** to output explicitly the graded set containing all pairs $(x, \mu_{A2}(x))$, for each x
- Compute $\mu_{A1 \wedge A2}(x) = \min\{\mu_{A1}(x), \mu_{A2}(x)\}$ for every x
- Return the top k objects, breaking arbitrarily the ties.

• Note: The constant k plays a role only at the 4th step...

13

QUERY TOP-K: algoritmo "ingenuo"

- per ogni elemento restituisci lista ordinata degli oggetti che soddisfano la prima subquery
- per ogni elemento restituisci lista ordinata degli oggetti che soddisfano la seconda subquery
- per ogni elemento calcola la funzione combinata
 - ordina i risultati e restituisci i primi k

---> la cardinalità dell'insieme voluto entra in gioco solo alla fine, quindi si calcola un risultato molto più grande di quello che serve in realtà --> enorme spreco di lavoro:
infatti è molto probabile che i k risultati che verranno restituiti siano stati "visitati" all'inizio

...improving efficiency... (Fagin)

- Assumption:
 - Sorted access
 - the subsystems can **output** the graded set consisting of all objects **one by one**, in sorted order based on grade, until the middleware system tells the subsystem to stop
 - the middleware system can later tell the subsystem to **resume** outputting the graded set where it left off.
 - (alternatively, the sorted objects can be output in sets of cardinality k , for any constant k , instead of one by one)
 - Random access
 - the middleware could ask the subsystem the grade, wrt a query, of any given object

14

Fagin propone un miglioramento alla tecnica naïve:
due tecniche di accesso ai dati

- SORTED ACCESS: possibilità di restituire per ciascuna query atomica la lista ordinata degli oggetti
- RANDOM ACCESS: data una query atomica e dato un oggetto, restituisce quanto l'oggetto soddisfa la query

Evaluating conjunctions of atomic queries

- Query: $Q = F_t(A_1, A_2, \dots, A_m)$, where
 - Q is monotonic
 - t is the aggregation function
 - each A_i is an atomic query, evaluated by the subsystem i .
- Assumption: there are at least k objects (so that "top k answers" make sense)

15

- query Q : combinazione di m query atomiche
 - **funzione di aggregazione monotona**
- m sottosistemi che valutano query atomiche

Algorithm

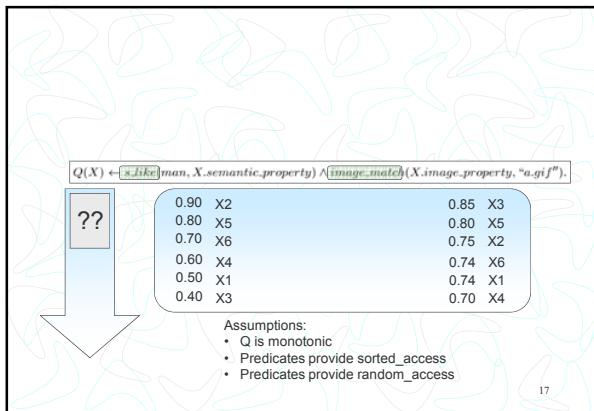
- **Sorted Access Phase:** for each i , give the query A_i under sorted access. Stop when there are at least k matches, that is, when there is a set L of at least k objects such that all the subsystems have output all the members of L
- **Random Access Phase:** for each object x that has been seen, do random access to each subsystem j to find $\mu_{A_j}(x)$
- **Computation Phase:** Compute the grade $\mu_Q(x) = \min(\mu_{A_1}(x), \mu_{A_2}(x), \dots, \mu_{A_k}(x))$ for each object x that has been seen.
- **Return** the graded set Y containing the objects with the k highest grades

16

- Fase di Sorted Access: si processano parallelamente le sottoquery ordinandole per grado di importanza.

- Fase di Random Access (accesso diretto): è possibile accedere ad un elemento restituito da una sottoquery ed analizzarne il valore.

- Fase di computazione: si combinano tutti i risultati parziali calcolando per ogni coppia di oggetti (nel caso di due sottoquery) il loro grado di soddisfacimento.

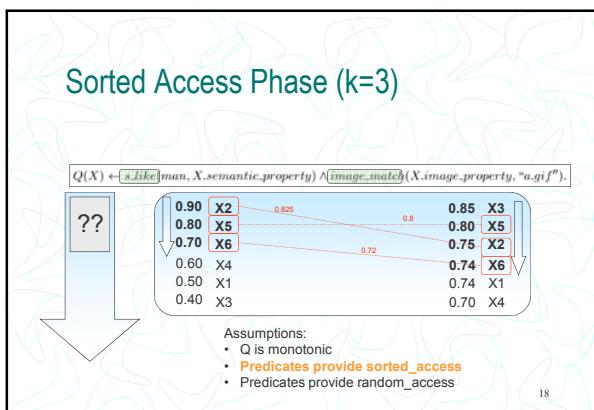


17

Ipotesi di lavoro

1) La funzione Q , ossia la query (ad esempio per A and B or C la funzione Q è $\min\{\max\{\mu_A(A), \mu_B(B)\}, C\}$) è monotona.

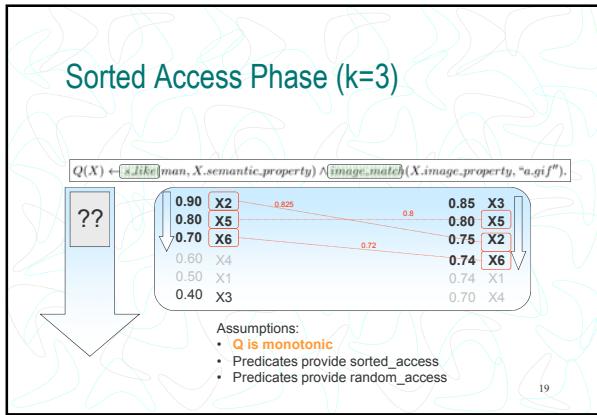
Sorted Access: partono due sequenze parallele di accessi ordinati, uno per la prima query e uno per la seconda. Accesso ordinato significa che si leggono i dati in modo ordinato (in base al valore di confidenza, ossia 0.90, 0.80, ecc. per la prima, 0.85, 0.80 ecc per la seconda).



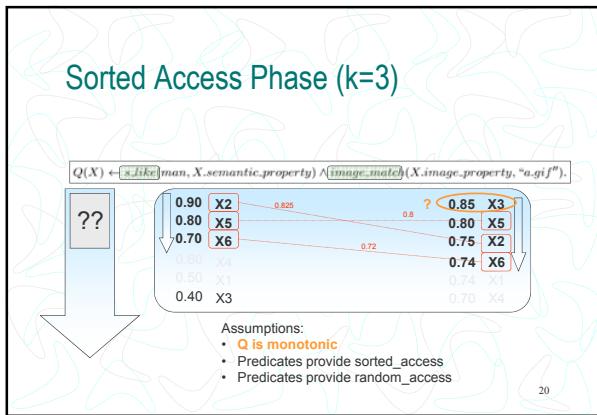
In questo caso mi fermo quando leggo X6 nella seconda query perché ho raggiunto $k=3$ risultati in comune tra le due sottoquery.

Avendone tre in comune, non posso concludere che questi tre in comune siano i migliori ma posso concludere di aver visto un insieme di elementi che include i top-3 perché Q è monotona, quindi per altri valori che sarebbero minori di questi (per il sorted access) per cui Q sarebbe minori.

Infatti i possibili candidati per la query sono i tre elementi in comune finora trovati più. Ma X3 è un altro possibile candidato perché non si ha la certezza che $Q(X_3)$ (prima query, X_3 seconda query) sia minore di Q che lega X1, X5 e X6 se come X3 ha un valore grande rispetto agli altri valori della query. Tuttavia ogni altro elemento in comune alle due sottoquery ha sicuramente Q minore di queste in quanto in entrambe le query tali elementi precedono questi nell'ordinamento (per cui anche Q è minore essendo monotona).

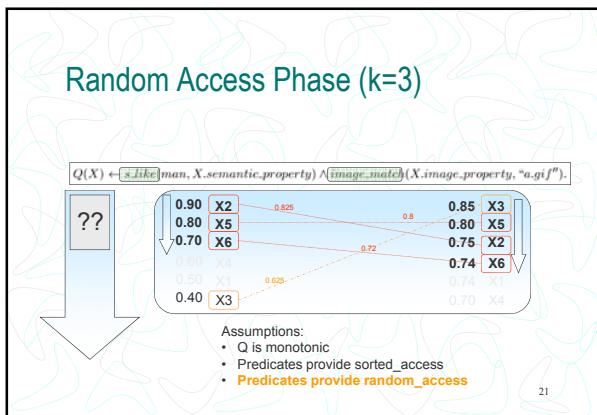


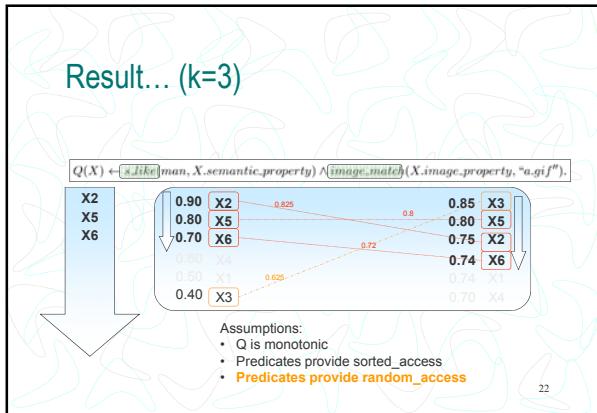
- n sottoquery indipendenti



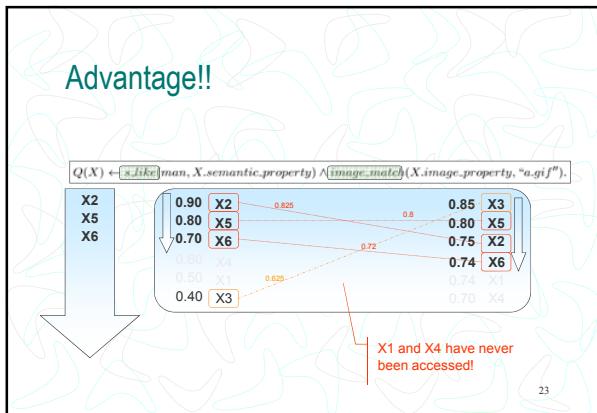
la fase di random access è importante!!!

infatti se X3 avesse avuto per la prima subquery valore 0.65, allora la media con 0.85 sarebbe stata di 0.75, e quindi X3 sarebbe entrato tra i primi k risultati





VANTAGGIO: abbiamo risparmiato l'accesso a X1 e X4!

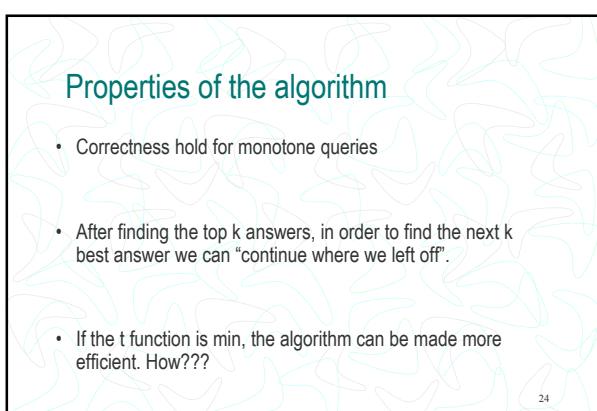


- se le query sono monotone l'algoritmo restituisce il risultato corretto: la monotonicità serve per poter terminare la fase di sorted access quando ho k elementi in comune, perché qualunque elemento non ancora visto è peggio rispetto a tutti i criteri (e quindi anche la sua combinazione!)

- consente comportamento incrementale: posso incrementare k e ricominciare da dove avevo interrotto

- se la funzione di aggregazione è il minimo, l'algoritmo si può rendere più efficiente sfruttando proprietà di

- * monotonicità
- * minimo



Use only one pred. for sorted access (k=3)

$Q(X) \leftarrow [x.\text{like}]man, X.\text{semantic_property} \wedge [image.\text{match}](X.\text{image_property}, "a.gif").$

??	0.90 X2 0.80 X5 ↓ 0.70 X6 0.60 X4 0.50 X1 0.40 X3	0.85 X3 0.80 X5 0.75 X2 0.74 X6 0.74 X1 0.70 X4
----	--	--

25

se uso il minimo, il risultato finale che restituisco è uno dei risultati che ho visto (non un valore creato ad hoc come la media) ---> allora posso fare sorted access solo su un predicato e poi, per ogni elemento, faccio random access dall'altra parte

Sorted+Random Access (k=3)

$Q(X) \leftarrow [x.\text{like}]man, X.\text{semantic_property} \wedge [image.\text{match}](X.\text{image_property}, "a.gif").$

??	0.90 X2 0.80 X5 ↓ 0.70 X6 0.60 X4 0.50 X1 0.40 X3	0.85 X3 0.80 X5 0.75 X2 0.74 X6 0.74 X1 0.70 X4
----	--	--

- Stop when the next value is smaller than the third candidate

26

posso fermarmi perché qualunque altro oggetto non può avere un minimo > 0.60!

---> ho sicuramente già visto i 3 minimi più alti

Sorted+Random Access (k=3)

$Q(X) \leftarrow [x.\text{like}]man, X.\text{semantic_property} \wedge [image.\text{match}](X.\text{image_property}, "a.gif").$

X2 X5 X6	0.90 X2 0.80 X5 ↓ 0.70 X6 0.60 X4 0.50 X1 0.40 X3	0.85 X3 0.80 X5 0.75 X2 0.74 X6 0.74 X1 0.70 X4
----------------	--	--

X1, X3, and X4 have never been accessed!

27

rispetto a prima ho evitato l'accesso anche ad X3

Evaluating disjunctions of atomic queries

- Query: $Q = A_1 \vee A_2 \vee \dots \vee A_m$
 - Q : standard fuzzy disjunction
- **Sorted Access Phase:** for each i , give the query A_i and collect the top k answers to this query
- **Computation Phase:** for each object x that has been returned by any of the m subsystems, compute $h(x) = \max\{\mu_{A_i}(x)\}$, for all i .
- **Return** the graded set Y containing the objects with the k highest grades $h(x)$

28

per la disgiunzione è più facile perché uso il MASSIMO:
- il max di due valori è uno dei due valori che ho visto

1) fase di SORTED ACCESS: mi fermo dopo aver preso i primi k oggetti per ogni subquery ---> siccome uso il massimo, per ogni sottosistema gli oggetti dal $k+1$ -esimo in poi che non sono stati visti in nessun'altra lista hanno sicuramente un massimo che è minore del più piccolo massimo trovato fino a quel momento --> la fase di random access non è necessaria!

Weighting the Importance of Subqueries

- $Q = (\text{AlbumColor} = \text{'black'}) \text{ and } (\text{Shape} = \text{'circle'})$
- What if the user cares twice as much about colors as shape?
 - Intuitively, we wish to assign twice as much weight to color as to shape
 - in the user interface, *sliders* are one mechanism to convey information about the weights
 - need to define the **weighted version of aggregation functions**

29

Weighting the Importance of Subqueries (ctd)

- The query Q is $A_1 \wedge A_2 \wedge \dots \wedge A_m$.
- g_i is the score of conjunct A_i
- f : function whose domain is the set of all tuples, of all sizes, over $[0,1]$, and with range $[0,1]$
- $f(g_1, \dots, g_m)$ is the overall score.
- $\theta_1, \dots, \theta_m \geq 0, \quad \sum \theta_i = 1 \quad \theta_i$ is the weight of attribute i
- $\Theta = (\theta_1, \dots, \theta_m)$ is a **weighting**

30

Weighting the Importance of Subqueries (ctd)

- For each weighting $\Theta = (\theta_1, \dots, \theta_m)$, a function f_Θ is derived, whose domain is the set of m-tuples
 - $f_\Theta(g_1, \dots, g_m)$ is the overall score, when the weights are given by the weighting Θ
 - if $\theta_1 \geq \theta_2 \geq \dots \geq \theta_m$, then $\Theta = (\theta_1, \dots, \theta_m)$ is said ordered.

31

Desiderata...

- $f_{(1/m, \dots, 1/m)}(g_1, \dots, g_m) = f(g_1, \dots, g_m)$
 - if all the weights are equal, the weighted function f_{Θ} coincides with the corresponding "unweighted" one
 - $f_{(\theta_1, \dots, \theta_{m-1}, 0)}(g_1, \dots, g_m) = f_{(\theta_1, \dots, \theta_{m-1})}(g_1, \dots, g_{m-1})$
 - if a particular argument has 0 weight, it can be dropped
 - $f_{(\theta_1, \dots, \theta_m)}(g_1, \dots, g_m)$ is a continuous function of $\theta_1, \dots, \theta_m$

32

Weighting formula

When $\theta_1 \geq \theta_2 \geq \dots \geq \theta_m$,

$$f_{\Theta}(g_1, \dots, g_m) = (\Theta_1 - \Theta_2) \cdot f(g_1) + \\ 2 \cdot (\Theta_2 - \Theta_3) \cdot f(g_1, g_2) + \\ 3 \cdot (\Theta_3 - \Theta_4) \cdot f(g_1, g_2, g_3) + \\ \dots + \\ m \cdot \Theta_m \cdot f(g_1, \dots, g_m)$$

Note: the weighting formula is well defined even when some θ_i are equal.

33

con una funzione di combinazione che soddisfa i desiderata di Fagin l'algoritmo funziona, perché l'hp alla base della correttezza dell'algoritmo è che la funzione sia monotona e continua (stessi assunti per i desiderata di Fagin)

Weighting formula (ctd)

- Monotonicity and strictness of the (unweighted) f is inherited by the weighted functions f_{Θ} .
- In particular, the optimal algorithm to evaluate conjunctions of atomic queries preserves correctness and optimality also in the weighted case.

34

Adding filter conditions (Chauduri and Gravano)

35

obiettivo: esprimere le query multimediali in modo simile a quelle relazionali aggiungendo condizioni di filtro tipo le clausole WHERE del modello relazionale

Query model requirements

- Account for the grade of match between the value of an attribute of a multimedia object, and a given constant

– Grade (attr, value) (o)

Real number
in [0,1]

attribute

object

36

- query atomica: "attributo = valore"

- Grade: grado di soddisfacimento della query atomica
"attr = value" da parte dell'oggetto o --> restituisce valore [0,1]

Query model requirements (2)

- Allow the user to specify thresholds on the grade of match of the acceptable objects (**filter conditions**)
 - Atomic filter condition:
 - Grade (attr, value) (o)>=grade
 - Compound conditions
 - conjunction/disjunction of conditions

37

posso aggiungere condizioni di filtro per restringere la ricerca

prima la ricerca era del tipo "cerca tutti gli oggetti che soddisfano al meglio tale condizione": qua invece vogliamo dire "cerca tutti gli oggetti che soddisfano questa condizione con grado superiore di un certo valore"

---> limito il sottoinsieme di dati a cui sono interessato: uso alcune condizioni solo come FILTRO per restringere il bacino di interesse

Query model requirements(3):

- Enable the user to ask for only a few top-matching objects (**ranking expressions**)
 - to compute a **composite grade** for an object from individual grades of match and the composition functions Min, Max

38

ora possiamo esprimere le condizioni secondo cui fare l'ordinamento (la condizione di filtro ora non interviene più!) e quindi si può usare ad es. l'algoritmo di Fagin

SQL-like syntax

- SQL-like syntax
 - Select oid
from Repository
where Filter_condition
Order[k] by Ranking_expression

39

...example

- The repository contains information about criminals
- A record on every person on file consists of
 - A textual description p (profile)
 - A scanned fingerprint f
 - A recording of a voice sample v.
- Query:**
 - Select oid
 - from Repository
 - where (Grade (v, V)>=0.5 and Grade(p,'american citizen')>=0.9)
or Grade(f,F) >=0.9
 - order[10] by max (grade(f,F), Grade(v,V))

40

- clausola WHERE: condizione di filtro

- clausola ORDER: condizione di ranking --> ordina in questo caso per valore massimo di grade(f,F) e grade (v,V)

utile perché il massimo cancella i valori delle singole subquery , quindi in questo modo possiamo impostare una soglia per considerare le singole subquery!

Expressivity of the query model

- Do we really need both filter condition (F) and ranking condition (R) ???
- Can we embed the filter condition in a new ranking expression R_F so that the top objects of R_F are the top objects for R that satisfy F ?

41

Le condizioni di filtro (ossia $\text{Grade}(x,X) \geq r$) aggiungono potere espressivo al linguaggio? SI.

Expressivity of the query model

- Do we really need both filter condition (F) and ranking condition (R) ???
- Can we embed the filter condition in a new ranking expression R_F so that the top objects of R_F are the top objects for R that satisfy F ?
 - NO (counterexample!!!!)

42

...counterexample

- Let $e1 = \text{Grade}(a1, v1)$
- $e2 = \text{Grade}(a2, v2)$
- Filter condition $F = e1 \geq 0.2$
- Ranking expression $R = e2$
- By contradiction, exists RF that satisfies both F and R
- RF equivalent to one of the following:
 $e1, e2, \min(e1, e2), \max(e1, e2)$

43

---> assumiamo per assurdo che la condizione di filtro possa essere incorporata nella condizione di ranking e troviamo un controsenso

- abbiamo due espressioni $e1$ ed $e2$
- usiamo come condizione di filtro $R = e1 > 0.2$
e condizione di ranking $R = e2$ (ordinamento in base a $e2$)

- supponiamo che esista un'espressione di ranking RF che soddisfi sia R che F

in questo caso RF può essere solo una di queste:
 $e1, e2, \min(e1, e2), \max(e1, e2)$

...database (example)

object	e1	e2	$\min(e1, e2)$	$\max(e1, e2)$
o1	0.1	0.6	0.1	0.6
o2	0.2	0.4	0.2	0.4
o3	0.5	0.3	0.3	0.5

44

dimostriamo che nessuno di questi RF restituirebbe lo stesso risultato di F seguito da R

- F seguito da $R \rightarrow o2 (0.4)$

mentre

- $RF = e1 \rightarrow o3 (0.5)$
- $RF = e2 \rightarrow o1 (0.6)$
- $RF = \min\{e1, e2\} \rightarrow o3 (0.3)$
- $RF = \max\{e1, e2\} \rightarrow o1 (0.6)$

qualunque delle 4 alternative lecite restituisce un risultato che è diverso da quello che otterrei con F ed R : quindi le condizioni di filtro aggiungono espressività

Storage-level access interface

- $\text{GradeSearch}(\text{attribute}, \text{value}, \text{min_grade})$
- $\text{TopSearch}(\text{attribute}, \text{value}, \text{count})$
- $\text{Probe}(\text{attribute}, \text{value}, \{\text{oid}\})$
- Not all the repositories have to support all of these interfaces at the physical level.

45

ipotesi di lavoro: abbiamo tre meccanismi a disposizione

1) GradeSearch è un tipo di accesso che restituisce tutti gli oggetti del DB che soddisfano con grado almeno min_grade la condizione atomica "attribute = value" ---> query di range

2) TopSearch restituisce i migliori count elementi rispetto alla condizione atomica "attribute = value" ---> query top-k

3) Probe (Random Access), dato un insieme di identificatori di oggetti e un attributo, restituisce il valore di soddisfacimento dell'attributo da parte di ogni oggetto

Filter conditions

- Assumption: the filter conditions are independent
A filter condition f is independent if:
 - Every atomic filter condition occurs at most once in f
 - $p(e_1 \wedge \dots \wedge e_n) = \prod_{i=1, n} p(e_i)$, being any e_i an atomic filter condition, and $p(e_i)$ the probability that the filter condition e_i is true.
- The repository requires the use of an index to evaluate every filter condition.

46

assumiamo anche che le condizioni di filtro siano indipendenti tra loro, ovvero che le condizioni atomiche siano

- ciascuna presente una volta sola, es. non c'è $A \text{ AND } A$
- soddisfano l'indipendenza a livello probabilistico, ovvero la P che siano soddisfatte tutte quante è pari al prodotto delle P che sia soddisfatta ciascuna condizione

inoltre abbiamo un indice per ogni attributo su cui abbiamo una condizione di filtro

Possible evaluation strategies

- Use one GRADESEARCH for EACH atomic condition in the filter condition, and then merge the returned sets of object ids through a sequence of union/intersections
- Use GRADESEARCH only for SOME atomic condition, and for the rest of the conditions use PROBE
- The key optimization problem becomes to determine the set of filter conditions that are evaluated using Gradesearch.

47

vogliamo ridurre il costo di valutazione della query globale

1) metodo ingenuo: per ciascuna condizione atomica invoca GradeSearch --> se la condizione era un and faccio l'intersezione degli insiemi, altrimenti se è una disgiunzione faccio l'unione ---> è molto costoso

2) individuo un sottoinsieme opportuno di condizioni atomiche su cui applicare il GradeSearch per ottenere un insieme su cui fare Probe per verificare il soddisfacimento delle altre condizioni ---> bisogna capire come individuare questo sottoinsieme di predicati atomici che, se soddisfatti, ci permettano di avere un sovrainsieme dei risultati finali (perché va bene se abbiamo false hits, ma non misses!)

Search minimal condition

- Goal: given a filter condition f , characterize the smallest sets of atomic conditions such that by searching the conditions in any of these sets we retrieve all the objects that satisfy f
- Example

$$f = a_4 \wedge ((a_1 \wedge a_2) \vee a_3)$$

48

1) con tecnica naïve:

- faccio GradeSearch su a_1, a_2, a_3, a_4
- ottengo insiemi di oggetti S_1, S_2, S_3, S_4
- calcolo insieme dei risultati come: $S_4 \cap ((S_1 \cap S_2) \cup S_3)$

2) vogliamo invece individuare un sottoinsieme di condizioni atomiche su cui fare GradeSearch e poi fare Probe sugli elementi dell'insieme restituito

Search minimal condition

- By searching on a condition using Grade Search we obtain a set of objects
- Some additional probes might be needed, to determine the set of objects that satisfy the rest of the condition as well
- → residue of f for a $\text{Res}(a, f)$

Boolean condition that the object retrieved using a must satisfy, in order to satisfy f.

49

- sulla SMC facciamo la GradeSearch

- su ciascun oggetto restituito in questo modo facciamo Probe per eliminare i false hits

bisogna verificare che l'oggetto effettivamente faccia parte del risultato: le verifiche sono espresse dal RESIDUO della condizione di filtro, ovvero la condizione che deve ancora essere verificata per far sì che un oggetto possa appartenere al risultato

...residue... examples

$$f = a_4 \wedge ((a_1 \wedge a_2) \vee a_3)$$

$$\text{Res}(a_2, f) = a_1 \wedge a_4$$

$$\text{Res}(a_4, f) = (a_1 \wedge a_2) \vee a_3$$

50

- se ho già valutato a2, allora per concludere che un oggetto soddisfi f è necessario che

- l'oggetto soddisfi a2 (è un AND)
- non è necessario che soddisfi a3 (è un OR)
- deve invece soddisfare anche a4 (è un AND)

come calcolare il residuo?

1) disegno l'albero sintattico della formula logica

2) parto dal nodo a2 e vado fino alla radice

3) aggiungo al residuo tutti i sottoalberi di nodi AND che non contengono a2

Search minimal conditions

GOAL: to characterize the **smallest sets** of atomic conditions such that, by searching the conditions in any of these sets, we retrieve all the objects that satisfy f (plus some extra ones that are pruned out with probing)

Example $f = a_4 \wedge ((a_1 \wedge a_2) \vee a_3)$

Search minimal condition sets:

$\{a_4\}$	$\{a_1, a_3\}$	$\{a_2, a_3\}$
-----------	----------------	----------------

51

Search Minimal Condition è un insieme minimale di condizioni che, se valutate, restituiscono un sovrainsieme del risultato finale (può contenere false hits, ma non misses)

---> non c'è nessun valore che non soddisfi gli insiemi SMC e che sia nel risultato

COME CALCOLO GLI INSIEMI SMC?
basta passare alla forma congiuntiva di f: questo mi dà la congiunzione degli insiemi SMC
 $f = a_4 \wedge (a_1 \vee a_3) \wedge (a_2 \vee a_3)$

Example (ctd)

- If we decide to search on $\{a_2, a_3\}$
 - Search on a_2 , and probe the retrieved objects with $\text{Res}(a_2, f) = a_1 \wedge a_4$.
Keep the objects that satisfy $\text{Res}(a_2, f)$
 - Search on a_3 , and probe the retrieved objects with $\text{Res}(a_3, f) = a_4$.
Keep the object that satisfy $\text{Res}(a_3, f)$
 - Return the objects kept.

52

supponiamo di scegliere $\{a_2, a_3\}$

1) faccio GradeSearch su a_2
e poi faccio Probe su $\text{Res}(a_2, f)$

2) faccio GradeSearch su a_3
e poi faccio Probe su $\text{Res}(a_3, f)$

3) restituisco l'unione degli elementi

Search minimal execution...

Execution like the one given in the previous example, based on any search minimal condition set.

53

Search minimal execution...

Execution like the one given in the previous example, based on any search minimal condition set.

How do we pick a plan from the space of search minimal executions?

54

come scegliamo il piano di esecuzione di una query?

The cost model

- Statistics associated with each atomic condition a:
 - Selectivity Factor $Sel(a)$
 - Fraction of the objects in the repository that satisfy the condition a
 - Search Cost $SC(a)$
 - Cost of retrieving the ids of the objects that satisfy the condition a using Grade Search
 - Probe Cost $PC(a,p)$
 - Cost of checking the condition a for p objects, using the probe access method.

55

si fa una stima dei costi per capire quali condizioni convenga anticipare (considerazioni statistiche che supponiamo di avere a disposizione):

- grado di selettività di una condizione a: percentuale di oggetti nel db che soddisfano a --> conviene anticipare un predicato altamente selettivo (per ridurre i Probe)
- costo di fare GradeSearch: costo di trovare gli oggetti che soddisfano una certa condizione sopra una certa soglia
- costo di fare Probe: costo di controllare la condizione a per p oggetti

Cost of the Search minimal executions

- m: search minimal condition, for the filter condition f
- w: algorithm for probing conditions
- $|O_a|$: number of objects that satisfy the condition a, that is, the product of the number of objects by the selectivity of a.
- $R(a,f)$: residue of f for a (boolean condition that the objects retrieved using a should satisfy to satisfy the entire condition f)
- $C_w(f,m) = \sum_{a \in m} (SC(a) + PC_w(R(a,f), |O_a|))$

56

il costo dell'esecuzione tiene conto del costo associato ai singoli passi effettuati

per ogni condizione atomica a nell'insieme SMC, si conta il costo di fare GradeSearch su a + il costo di fare Probe su una formula che è il residuo di a per tutti gli oggetti che soddisfano a

Optimal search minimal condition for

- Let f be a filter condition, and f' a subexpression of f
- Inductive search-minimal condition for f' wrt f:
 - If $f' = a$ (atomic condition), then $SM_i(f') = \{a\}$
 - If $f' = f_1 \wedge \dots \wedge f_n$, then $SM_i(f') = SM_i(f_1) \wedge \dots \wedge SM_i(f_n)$, where
 - $C(f, SM_i(f_i)) = \min \{ C(f, SM_i(f_1)), \dots, C(f, SM_i(f_n)) \}$
 - break the ties arbitrarily
 - If $f' = f_1 \vee \dots \vee f_n$, then $SM_i(f') = SM_i(f_1) \cup \dots \cup SM_i(f_n)$
- If f is independent, then $SM_i(f)$ is optimal

57

definizione induttiva:

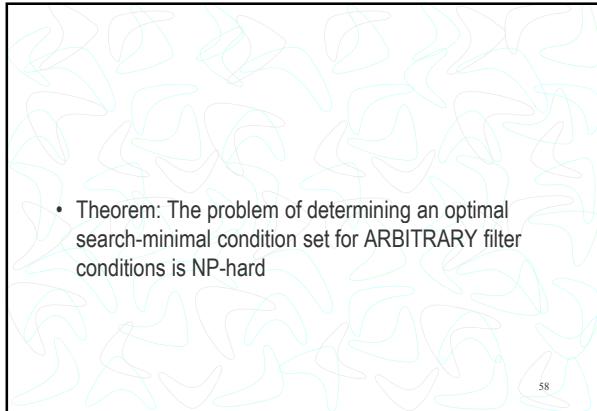
se f' è sottoespressione di f, $SMC(f')$ è:

- se f' è atomica, allora $f' = a$ e $SMC(f') = \{a\}$

- se $f' = f_1 \wedge \dots \wedge f_n$ devo individuare il congiunto ideale, che è quello scelto in modo tale da minimizzare il costo stimato: quindi $SMC = \{f_i\}$ tale che $C(f, SMC(f_i)) = \min \{ C(f, SMC(f_j)) \}$ per ogni j

- se $f' = f_1 \vee \dots \vee f_n$, allora $SMC(f') = SMC(f_1) \cup \dots \cup SMC(f_n)$

--> se f è indipendente, allora la SMC estratta è ottimale, ovvero ha costo minimo



- Theorem: The problem of determining an optimal search-minimal condition set for ARBITRARY filter conditions is NP-hard

58

se le condizioni atomiche non sono indipendenti allora il problema di determinare l'insieme SMC è NP-hard

Web

- A network of pages
 - very large
 - links carry information
- Keyword-based query
 - queries are underspecified
 - average 1-2 keywords

99
2

Maria Luisa Sapino (BDMM 2010)

Tutte le tecniche che abbiamo visto si applicano bene al web. Basti pensare a google.

Web

- Approach 1: use standard IR techniques to find pages that satisfy a query

99
3

Maria Luisa Sapino (BDMM 2010)

Come fare ricerca di informazione sul web?

1) approccio naïve: La query è fatta di parole chiave. Si effettua ricerca classica su testo (come abbiamo visto) ignorando i link.

Il problema è che il link portano semantica.

Web

- Approach 1: use standard IR techniques to find pages that satisfy a query

99
4

Maria Luisa Sapino (BDMM 2010)

problema: non riusciamo ad esprimere aspetti dinamici relativi all'organizzazione delle informazioni

è opportuno che la navigazione tra i diversi documenti segua un ordinamento logico, che esprima certe relazioni di riferimenti o dipendenze per consentire una lettura ordinata dei contenuti che sarebbe più efficace

---> non posso raggiungere questi obiettivi con tecniche di IR standard

99

5

Web

- Approach 1: use standard IR techniques to find pages that satisfy a query

Maria Luisa Sapino (BDMM 2010)

l'esistenza di link è portatrice di informazione anche se non espressa direttamente nella pagina

99

6

Web

- Approach 1: use standard IR techniques to find pages that satisfy a query

Maria Luisa Sapino (BDMM 2010)

in una pagina sono presenti tutte e 3 le keyword, nell'altra solo 2: una tecnica di IR standard scarterebbe questa pagina, anche se questa LINKA ad un'altra che invece parla di tutte e 3

99

7

Web

- Approach 2: integrate IR techniques with structure/link analysis

Maria Luisa Sapino (BDMM 2010)

2) sfruttiamo i link per non perdere queste informazioni: integriamo sistemi di IR con tecniche di link analysis

Web

- Approach 2: integrate IR techniques with structure/link analysis

(a) Connectivity	(b) Co-citation	(c) Social filtering	(d) Transitivity

Maria Luisa Sapino (BDMM 2010)

99
8

identifichiamo 4 relazioni che possono esistere tra due pagine diverse e che hanno connotazione sociale diversa:

- 1) CONNELLITIVITA': due pagine sono connesse se esiste un percorso da una all'altra in entrambi i sensi ---> possiamo vederlo allora come un documento unico
- 2) CO-CITAZIONE: stesso documento che linka due documenti
- 3) SOCIAL FILTERING: i due documenti puntano ad una stessa pagina ---> dichiarazione di importanza che è riferimento di molte altre pagine
- 4) TRANSITIVITA': raggiungo un documento dall'altro tramite un documento

HITS algorithm

- Good pages are categorized into two types
 - Hubs: point to many pages of high quality
 - Authorities: pages of high quality

Maria Luisa Sapino (BDMM 2010)

99
9

due categorie di pagine (ognuna può essere sia hub che authority):

- HUB: punto di smistamento, cioè una pagina che punta a tante altre pagine ---> un hub è tanto migliore quanto più sono di impatto le pagine a cui lui punta
- AUTHORITY: pagina di alta qualità, ovvero che è puntata da tante pagine

Hubs and authorities

Maria Luisa Sapino (BDMM 2010)

10
00

Hubs and authorities

10
01

Maria Luisa Sapino (BDMM 2010)

la qualità del link è direttamente proporzionale alla qualità della destinazione, e viceversa

Topic distillation by iterative mutual reinforcement

10
02

Maria Luisa Sapino (BDMM 2010)

vogliamo individuare l'insieme di pagine che caratterizzano un certo topic:
dobbiamo quantificare il grado di autorevolezza rispetto al grado di hubness

autorevolezza di P è la somma del grado di hubness delle pagine che puntano a P

Topic distillation by iterative mutual reinforcement

10
03

Maria Luisa Sapino (BDMM 2010)

il grado di hubness di P è la somma dell'autorevolezza delle pagine puntate da P

HITS

- Use IR to find the candidate pages

10
04

Maria Luisa Sapino (BDMM 2010)

HITS

- Use IR to find the candidate pages
- Expand to include all pages which link or are linked by this core set

10
05

Maria Luisa Sapino (BDMM 2010)

- Use IR to find the candidate pages
 - Expand to include all pages which link or are linked by this core set

Multidisciplinary (PRMM 0010)

HITS

$$Gq = \langle Vq, Eq \rangle$$

- Use IR to find the candidate pages
- Expand to include all pages which link or are linked by this core set
- Compute authority and hub values for all pages (iterate!!)

$$a(i) = \sum_{j \in in(i)} h(j) \quad h(i) = \sum_{j \in out(i)} a(j)$$

$$a(i) = \sum_{j \in in(i)} h(j) \quad h(i) = \sum_{j \in out(i)} a(j)$$

- Finally, HITS solves these mathematical equations to identify hubs and authority scores of the pages in V_q and selects those pages with high authority scores to be presented to the user as answers to the query
(dal libro)

come individuare insieme di pagine che concorrono alla caratterizzazione di un certo argomento e che facciano emergere informazioni sulla base di hub e authorities

AI GORITMO HITS

- 1) a fronte di una query: "trovare tutte le pagine che parlano di DBM" vengono restituiti con IR standard le pagine che più matchano con la mia query; da queste 2) HITS espande il nucleo delle pagine restituite con quelle adiacenti (puntate o che sono puntate) 3) iterativamente calcola il grado di autorevolezza e hubnes di queste pagine, cioè aggiorno prima hubness e in base a quel valore aggiorno autorevolezza (ecco perché “iterativamente”)

dal libro: "Given a web search query, the HITS algorithm tries to locate good authorities related to the query to help prevent poor pages from being returned as results to the user."

HITS

- Matrix notation

$$\vec{a} = E^T \vec{h} \qquad \qquad \vec{h} = E \vec{a}$$

10
07

Maria Luisa Sapino (BDMM 2010)

questo calcolo iterativo si può fare con calcolo matriciale vedendo a e h come vettori e passando dalla matrice di connettività del grafo

- E è la matrice che rappresenta gli archi delle pagine in uscita da un nodo (la usiamo per calcolare l'hubness)

- E' è la matrice degli archi trasposta e che quindi rappresenta gli archi in entrata (serve per calcolare l'autorevolezza)

...reminder

- Eigenvalue and eigenvector
- Given a matrix E , let c (scalar) and x (vector) be such that

$$c \xrightarrow{\quad} x = E \xrightarrow{\quad} x$$

 Eigenvalue Eigenvector

10
08

Maria Luisa Sapino (BDMM 2010)

...authorities

10
10

...authorities

$$\vec{a} = E^T E \vec{a}$$

a is an eigenvector of $E^T E$

Maria Luisa Sapino (BDMM 2010)

sapendo che
 $\vec{h} = E \cdot \vec{a}$

 posso riscrivere $\vec{a} = E^T \cdot \vec{h}$ come
 $\vec{a} = E^T \cdot E \cdot \vec{a}$

 e vedere
 - \vec{a} come autovettore di $E^T \cdot E$
 - 1 come autovalore di \vec{a}

10
11

...hubs

$$\vec{h} = E E^T \vec{h}$$

h is an eigenvector of $E E^T$

Maria Luisa Sapino (BDMM 2010)

analogamente per l'hubness

10
12

HITS and LSI???

- ...reminder: SVD of E is

$$E = U M V^T$$

where

- M is a diagonal matrix
- $U^T U = I$
- $V^T V = I$

Maria Luisa Sapino (BDMM 2010)

In questo modo però si ottengono false hits, in particolare questo avviene se una pagina rilevante (ossia pertinente alla query effettuata) punta (tramite link) ad una molto autorevole che però non è correlata con quello che si intende cercare.

Idea: pesare ogni arco del grafo (hub,authority) in termini di contributo, ossia un arco $a \rightarrow b$ è pesato w se a collega b dove la pagina b ha attinenza w . Quindi i gradi di autorità e hub sono calcolate in base alla query effettuate e quindi variano ad ogni query.

Questo significa però che per ogni query bisogna calcolare i valori di autorità ed host, mentre con la soluzione precedente (grafo non pesato), assumendo che i documenti non vengono aggiornati frequentemente, si possono memorizzare nel database tutti i valori di autorità ed host una volta sola (ed eventualmente aggiornarli una volta a settimana ad esempio, ma NON ad ogni query e comunque non è necessario tenere conto di autorità e host per ogni query).

Questo significa che questo metodo (HITS and LSI) è buono dal punto di vista teorico ma difficilmente applicabile per questioni di efficientia.

si può scomporre E in un prodotto di tre matrici tali che:
 - M è una matrice diagonale (ha valori diversi da 0 solo sulla diagonale)

- $U^T \cdot U = I$
- $V^T \cdot V = I$

..then

$$EE^T = UMV^T VMU^T$$

10
13

Maria Luisa Sapino (BDMM 2010)

se $E = UMV'$ (dove V' è la trasposta di V)

e ricordando che, date due matrici A e B ,
 $(A \cdot B)' = B' \cdot A'$

$$E' = (U \cdot M \cdot V')' = (V')' \cdot M' \cdot U' = V \cdot M' \cdot U'$$

ma siccome M è una matrice diagonale, $M = M'$, quindi
 $E' = V \cdot M \cdot U'$

..then

$$EE^T = UMV^T VMU^T \quad \text{ma } V' \cdot V = I$$

$$EE^T = UMIMU^T \quad \text{ma } M \cdot I = M$$

$$EE^T = UM^2U^T$$

10
14

Maria Luisa Sapino (BDMM 2010)

...then again

$$EE^T = UM^2U^T$$

$$EE^T U = UM^2U^T U = UM^2$$

---> perché $U' \cdot U = I$

10
15

Maria Luisa Sapino (BDMM 2010)

...then again

$$EE^T = UM^2U^T$$

$$EE^TU = UM^2U^TU = UM^2$$

$$EE^TU_j = U_j m_j^2$$

Eigenvector Eigenvalue

Maria Luisa Sapino (BDMM 2010)

10
16

ma M^2 è la matrice diagonale M in cui tutti gli elementi sulla diagonale sono elevati al quadrato --> quindi è ancora una matrice diagonale

di conseguenza possiamo vedere l'equazione come

$$E \cdot E^T \cdot U_j = U_j \cdot m_j^2$$

dove m_j è il j-esimo elemento della matrice M
e U_j è la j-esima riga di U

...in fact

$$EE^TU_j = U_j m_j^2$$

Eigenvector Eigenvalue

$$EE^T \vec{h} = \vec{h}$$

$\vec{h} = U_j$

10
17

ma allora $\vec{h} = U_j$ per qualsiasi j

...in fact

$$EE^TU_j = U_j m_j^2$$

Eigenvector Eigenvalue

$$EE^T \vec{h} = \vec{h}$$

$\vec{h} = U_j$

10
18

HITS is similar to LSI, but on (source, destination) rather than (term, document) matrix

Maria Luisa Sapino (BDMM 2010)

quindi il comportamento dell'algoritmo HITS è analogo a quello applicato per SVD e quindi possiamo lavorare non iterativamente ma sfruttando il calcolo matriciale

dal libro: "Note that a potential problem with the direct application of the foregoing technique for web search is that, although the relevant web neighborhood (G_q) is identified using the query, q , the neighborhood also contains pages that are not necessarily relevant to the query, and it is possible that one of these pages will be identified as the highest authority in the neighborhood. This problem, where authoritative pages are returned as results even if they are not directly relevant to the query, is referred to as topic drift. Such topic drift can be avoided by considering the content of the pages in addition to the links in the definition of hubs and authorities."

Problem

- Topic drift
 - Pages include neighbors
 - Neighbors may be good hubs, authorities; but may not have good content match

10
19

Maria Luisa Sapino (BDMM 2010)

PROBLEMA: non è detto che i nodi adiacenti solo perché con alto hubness o autorità siano risultati desiderati

ad es. ho nella mia pagina i link alle pagine dei miei familiari che sono su facebook --> allora sono molto autorevoli, però non voglio restituirli

Clever system

- Use IR to find the candidate pages
- Expand to include all pages which link or are linked by this core set
- Compute authority and hub values for all pages (iterate!!)
- Consider text next to the link!!!!

$$a(i) = \sum_{j \in in(i)} w(j \rightarrow i) h(j) \quad h(i) = \sum_{j \in out(i)} w(i \rightarrow j) a(j)$$

10
20

Maria Luisa Sapino (BDMM 2010)

MIGLIORAMENTO:
aggiungo una fase all'algoritmo HITS, tenendo conto del testo che circonda il link

ad es. se vicino al mio link c'è scritto "i miei familiari" e sono in un articolo scientifico, allora si associa un peso basso a quel link e non viene restituito

Problem

- Topic drift
 - Pages include neighbors
 - Neighbors may be good hubs, authorities; but may not have good content match
- Slow
 - Iteration is not good!
 - One eigenvector computation per query

10
21

Maria Luisa Sapino (BDMM 2010)

PROBLEMA:

- questo non ci consente di utilizzare il calcolo matriciale: il meccanismo iterativo è molto costoso con l'introduzione dei pesi w

dal libro: "In order to avoid query-time link analysis, the PageRank algorithm performs the link analysis as an offline process independently of the query. Thus, the entire web is analyzed and each web page is assigned a pagerank score denoting how important the page is based on structural evidence. At the query time, the keyword scores of the pages are combined with the pagerank scores to identify the best matches by content and structure."

PageRank

- Random Surfer
 - Jumps from page to page with uniform probability
 - Occasionally jump to a random page with small probability ($1-\beta$)
 - If no out page, then jump to any page with equal probability

10
22

Maria Luisa Sapino (BDMM 2010)

un utente casuale (Surfer) si trova in una pagina i e può passare ad un'altra pagina j

- o seguendo il link all'interno della pagina
- o per qualunque altro motivo che non possiamo conoscere

il salto ad un'altra pagina è disciplinato dal numero di link all'interno della pagina:

- la probabilità di saltare da una pagina ad un'altra è uguale per tutti i link (uniforme) all'interno di una stessa pagina, ad es. se ci sono 18 link allora $P = 1/18$

- la probabilità di saltare ad una pagina random per qualunque motivo è $(1-\beta)$, dove β è la probabilità di restare all'interno della pagina corrente

PageRank

- Random Surfer (N pages)
 - Jumps from page to page with uniform probability
 - Occasionally jump to a random page with small probability ($1-\beta$)
 - If no out page, then jump to any page with equal probability

$$\mathbf{Z} = (1 - \beta) \left[\frac{1}{N} \right]_{N \times N} + \beta \mathbf{M}$$

Transition matrix

$M_{ij} = \begin{cases} \frac{1}{out(i)} & \text{if there is an edge from } i \text{ to } j \\ 0 & \text{otherwise} \end{cases}$

10
23

Maria Luisa Sapino (BDMM 2010)

Z: matrice di transizione da una pagina all'altra

$(1-\beta)$: probabilità di un salto occasionale ad una qualunque altra pagina

$[1/N]$: è la matrice di transizione da una pagina ad un'altra, dove N è il numero di nodi del grafo e all'interno di ogni cella di questa matrice ciascuna pagina ha probabilità $(1/N)$ a cui salto con probabilità $(1-\beta)$

β : è la probabilità di restare all'interno della pagina corrente

M: la probabilità di continuare la navigazione seguendo un link all'interno della pagina corrente, ovvero una matrice in cui $M[i,j]$ è la probabilità di passare dalla pagina i alla pagina j seguendo un link all'interno della stessa pagina, cioè $1/($ numero di archi in uscita da i) (0 se non ci sono archi in uscita)

PageRank

- Random Surfer (N pages)
 - Jumps from page to page with uniform probability
 - Occasionally jump to a random page with small probability ($1-\beta$)
 - If no out page, then jump to any page with equal probability

$$P(j) = \frac{1-\beta}{N} + \beta \sum_{i \in in(j)} \frac{P(i)}{out(i)}$$

Probability that the surfer is at page j

10
24

Maria Luisa Sapino (BDMM 2010)

- $P(j)$ è la probabilità che l'utente qualunque sia nella pagina j , che coincide anche con l'autovettore della matrice di transizione Z

- $P(i)$ è la probabilità di seguire un link all'interno della pagina stessa

- $out(i)$ è il numero di archi in uscita

dal libro: "the components of the first eigenvector of Z will give the portion of the time spent at each node after an infinite run; that is, (similarly to HITS) the components of this eigenvector can be used as the pagerank scores of the pages (denoting how important the page is based on link evidence)."

PageRank

- Random Surfer (N pages)
 - Jumps from page to page with uniform probability
 - Occasionally jump to a random page with small probability ($1-\beta$)
 - If no out page, then jump to any page with equal probability

$$P(j) = \frac{1-\beta}{N} + \beta \sum_{i \in in(j)} \frac{P(i)}{out(i)}$$

Probability that the surfer is at page j
Primary eigenvector of the transition matrix Z

Maria Luisa Sapino (BDMM 2010)

PageRank

$$R(u) = \frac{1}{c} \sum_{v \in B_u} \frac{R(v)}{N_v}$$

Maria Luisa Sapino (BDMM 2010)

PageRank and Content

- Query independent
 - Query score has to be combined with PageRank score

Maria Luisa Sapino (BDMM 2010)

PageRank and Content

- Query independent
 - Query score has to be combined with PageRank score
- Topic Drift
 - ...maybe PageRank should be computed after the relevant community of pages are identified.

10
28

Maria Luisa Sapino (BDMM 2010)
