Sistemi di information retrieval e HCIR

Dott. Giuseppe Desolda

Outline

- Panoramica sui sistemi IR
- Progettazione di interfacce per i sistemi IR
- Framework di valutazione
- Tecniche di visualizzazione e ricerca delle informazioni

Information Retrieval (IR): definizione (1/2)

- “Obiettivo dell’IR è recuperare, all’interno di un insieme di documenti, tutti e solo i documenti rilevanti per un particolare utente con una particolare richiesta” (Calvin Mooers 1952)
- Rappresentazione, memorizzazione e organizzazione informazione non strutturata
  - Testi
  - Pagine web
  - Contenuti multimediali
- Scopo: trovare informazioni utili e rilevanti per l’utente.
- Esempi di bisogni da soddisfare

3

Information Retrieval (IR): definizione (2/2)

- Forti differenze rispetto alle base di dati
  - L’enfasi non è quindi sulla ricerca di dati
  - Si basa sulla ricerca di informazioni.
- Studiato fin dagli anni `70, negli anni `90 l’esplosione del Web ha moltiplicato l’interesse per IR.
- Nato per i dati strutturati, l’IR si é poi prestato bene a
  - Dati multimediali (immagini, audio, video)
  - Semi-strutturati (XML).
- Il problema è che non è semplice caratterizzare esattamente i bisogni informativi dell’utente.
  - Ambiguità
  - Complessità dei linguaggi naturali.
  - Domini d’uso differenti

4

IR e Data Retrieval a confronto

UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO

- Un sistema di Data Retrieval (ad esempio un DBMS) gestisce dati che hanno una struttura ben definita.
- Un sistema di Information Retrieval gestisce testi scritti in linguaggio naturale, spesso non ben strutturati e semanticamente ambigui.
- Di conseguenza:
  - Un linguaggio per Data Retrieval permette di trovare tutti gli oggetti che soddisfano esattamente le condizioni definite. Tali linguaggi (algebra relazionale, SQL) garantiscono una risposta corretta e completa e di manipolare la risposta.
  - Un sistema di IR, invece, potrebbe restituire anche oggetti non esatti (risultati non pertinenti); l'importante è che siano piccoli errori accettabili per l'utente.

5

Tecniche di IR: inverted index(1/4)

UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO

- I sistemi di IR operano su vista logica dei documenti
  - Non fattibile sui documenti originali
  - I documenti di una collezione vengono rappresentati tramite un insieme di keyword.
- I moderni elaboratori rappresentano un documento tramite l'intero insieme delle parole: vista logica full text.
  - Per collezioni molto grandi tale tecnica può essere inutilizzabile
  - Si utilizzano tecniche di modifica del testo per ridurre la dimensione della vista logica, che diventa un insieme di index term.
- Il modulo di gestione della collezione si occupa di creare gli opportuni indici, contenenti tali termini.

6

Tecniche di IR: inverted index (2/4)

UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO

- Le tecniche di indicizzazione studiate per le basi di dati relazionali (ad es. B-Tree) non sono adatte per i sistemi di Information Retrieval.
- L'indice più utilizzato dagli IR è l'indice invertito (inverted index):
  - Viene memorizzato l'elenco dei termini contenuti nei documenti della collezione
  - Per ogni termine, viene mantenuta una lista dei documenti nei quali tale termine compare.
- Tale tecnica è valida per query semplici (insiemi di termini)
  - Modifiche sono necessarie se si vogliono gestire altre tipologie di query (frasi, prossimità ecc.).

7

Tecniche di IR: inverted index (3/4)

UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO

- Il numero di termini indicizzati viene ridotto utilizzando una serie di tecniche, tra cui:
  - Eliminazione delle stopword: articoli, congiunzioni ecc.;
    - "Il cane di Luca" -> "cane Luca"
  - De-hyphenation: divisione di parole con trattino;
    - "Sotto-colonnello" -> Sotto colonnello
  - Stemming: riduzione alla radice grammaticale:
    - "Mangiano, mangiamo, mangiassi" -> "Mangiare"
  - Thesauri: gestione dei sinonimi, omonimi, ipernonimi
    - "Casa" -> "Casa, magione, abitazione, ..."
  - L'utilizzo di tali tecniche non sempre migliora la qualità delle risposte ad una query.

8

Tecniche di IR: inverted index (4/4)

- Avendo quindi non i documenti ma i loro inverted index, il processo di ricerca di informazioni viene riformulato:
  1. L'utente specifica un bisogno informativo.
  2. La query viene eventualmente trasformata...
  3. ...per poi essere eseguita, utilizzando indici precedentemente costruiti, al fine di trovare documenti rilevanti;
  4. I documenti trovati vengono ordinati in base alla (presunta) rilevanza e ritornati in tale ordine all'utente;
  5. L'utente esamina i documenti ritornati ed eventualmente raffina la query, dando il via ad un nuovo ciclo.

9

Modelli di IR: booleano

- E' il modello più semplice
  - Si basa sulla teoria degli insiemi e l'algebra booleana.
  - E' stato il primo ed il più utilizzato per decenni.
  - Oggi non viene quasi più considerato nei casi reali
- I documenti vengono rappresentati come insiemi di termini.
- Le query vengono specificate come espressioni booleane,
  - elenco di termini connessi dagli operatori booleani AND, OR e NOT.
- La strategia di ricerca è basata su un criterio di decisione binario, senza alcuna nozione di grado di rilevanza:
  - un documento è considerato rilevante oppure non rilevante.

10

Modelli per IR: vettoriale

- Assegnare un giudizio binario ai documenti (1=rilevante, 0=non rilevante del booleano) è troppo limitativo.
- Ad ogni termine nei documenti o nelle query viene assegnato un peso (un numero reale positivo).
- I documenti e le query sono rappresentati come vettori in uno spazio n-dimensionale (n = # di termini indicizzati).
- La ricerca viene svolta calcolando il grado di similarità tra
  - il vettore che rappresenta la query
  - i vettori che rappresentano ogni singolo documento
  - i documenti con più alto grado di similarità con la query hanno più probabilità di essere rilevanti per l'utente.
- Il grado di similarità viene quantificato utilizzando una misura
  - un esempio é il coseno dell'angolo tra i due vettori (che esprime la "vicinanza" fra i vettori).

11

Prestazioni di un sistema IR

- Come è possibile rispondere alla domanda "quale di questi due sistemi di IR funziona meglio"?
- Le prestazioni di un sistema tradizionale di Data Retrieval possono essere valutate oggettivamente, sulla base delle performance (velocità di indicizzazione, ricerca ecc.).
- In un sistema di IR tali valutazioni sono piú complesse
  - Causa: soggettività delle risposte alle query, domini differenti,…
  - Quello che si vorrebbe in qualche modo misurare è la "soddisfazione" dell'utente.
- Esistono delle misure standard per valutare la bontà delle risposte fornite da un sistema di IR: precision e recall sono un classico esempio

12

Prestazioni degli IR: precision e recall

UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO

Some classic IR evaluation measures.

Measure	Description
Recall	The number of retrieved relevant documents divided by the number of relevant documents in corpus.
Precision	The number of relevant retrieved documents divided by the number of retrieved documents.
F-measure	The F-measure is a way of combining precision and recall and is equal to their weighted harmonic mean $[F = 2(\text{precision} \times \text{recall})/(\text{precision} + \text{recall})]$. The F-measure also accommodates weighting of precision or recall, to indicate importance.
Average precision (AP)	Individual precision scores are computed for each relevant retrieved document (with 0 assigned to relevant documents that are not retrieved). These values are then summed and divided by the total number of relevant documents in the collection. Thus, AP has a recall component to it and is typically described as the area underneath the precision/recall curve. AP also takes into account the position of relevant documents in the result list.
Mean average precision (MAP)	This is a run level measure and consists of taking the average of the average precision values for each topic.
Geometric average precision (GMAP)	The geometric mean of n values is the nth root of the product of the n values. Robertson [219] recommends taking the logs of the values and then averaging. GMAP was developed for the TREC Robust Track, which explored retrieval for difficult topics and does a better job than MAP of distinguishing performance scores at the low end of the AP scale.
Precision at n	The number of relevant documents in the top n results divided by n. Typical values for n are 10 and 20, which is thought to better represent the user's experience since research has shown that this is the extent to which users look through Web search results [146].
Mean reciprocal rank (MRR)	This measure was developed for high-precision tasks where only one or a small number of relevant documents are needed. For a single task with one relevant document, reciprocal rank is the inverse of its ranked position. MRR is the average of two or more reciprocal rank scores (used when there is more than one task).

13

Prestazioni degli IR: precision e recall

UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO

Interactive Recall and Precision

Modified versions of recall and precision for interactive IR [284, 285] and relative relevance [33, 36].

Measure	Description
Interactive recall	Number of TREC relevant saved by user/number of TREC relevant documents in the corpus.
Interactive TREC precision	Number of TREC relevant documents viewed by the user/total number viewed.
Interactive user precision	Number of TREC relevant documents saved by the user/total number saved by the user.
Relative relevance (RR)	Cosine similarity measure between two lists of relevance assessments for the same documents.

14

Prestazioni degli IR: precision e recall

- In conclusione, l'utilizzo di Precision e Recall per la valutazione di un motore di IR pone alcuni problemi:
- I documenti della collezione devono essere valutati manualmente da persone esperte: non sempre il giudizio è completamente veritiero;
- La valutazione dei documenti è binaria (rilevante / non rilevante): non sempre è facile catalogare così nettamente un documento;
- Le misure sono pesantemente influenzate dal dominio di applicazione, cioè dalla collezione e dalle query: un motore di IR potrebbe avere delle ottime prestazioni in un determinato dominio ma non in un'altro.

15

IR per il Web

- L'IR è nata per gestire collezioni statiche e ben conosciute: testi di legge, enciclopedie ecc., ma quando la collezione di riferimento diventa il Web, le cose cambiano completamente:
- La collezione è dinamica, molto variabile nel tempo;
- Le dimensioni sono enormi;
- I documenti non sono sempre disponibili;
- Le query degli utenti sono ancora più imprecise e vaghe.
- Le tecniche utilizzate dai motori di ricerca possono quindi differire, ad esempio Pagerank (una cui variante è usata da Google) utilizza criteri diversi dalla IR standard

16

Advanced IR

- Per incrementare l'efficacia dell'IR, diverse tecniche sono utilizzate:
- Ranking probabilistico
- Latent semantic Indexing
- Relevance FeedBack e Query Expansion

17

Advanced IR: ranking probabilistico

- Il modello probabilistico: Il principio di pesatura probabilistico, o probability ranking principle
- Metodi di ranking:
  - Binary Independence Model
  - Bayesian networks
- L'idea chiave è di classificare i documenti in ordine di probabilità di rilevanza rispetto all'informazione richiesta: P(rilevante|documentoi, query)

18

Advanced IR: latent semantic indexing

UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO

- I metodi di ranking tradizionali calcolano l'attinenza di un documento ad una query sulla base della presenza o meno di parole contenute nella query: un termine o è presente o non lo è
- Nel LSI la ricerca avviene per concetti
  - un concetto non è l'astrazione-generalizzazione di un termine (es: *golf vestiario*)
  - È un insieme di termini correlati (golf, maglia, vestito) detti co-occorrenze o **dominio semantico**
- Si ottiene mediante la fattorizzazione SVD della matrice termini documenti

Advanced IR: Relevance FeedBack e Query Expansion

UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO

- Il Relevance Feedback e Query Expansion sono tecniche per migliorare il recall di una query.
- Nel *Relevance Feedback*
  1. Alla fine di ogni ricerca si presentano un set iniziale di documenti
  2. Si chiede all'utente di selezionare i più rilevanti (anche in maniera trasparente all'utente)
  3. Si usa questo feedback per riformulare la query
- Nella Query expansion
  - L'obiettivo è di migliorare la qualità della ricerca
  - Si aggiungono termini oltre quelli iniziali (es: sinonimi, termini più ricorrenti nei documenti ritrovati, stemming, acronimi, traduzioni, etc),
  - Il peso dei termini aggiunti è solitamente inferiore ai termini dell'utente

Progettazione e valutazione di interfacce per sistemi di IR

UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO

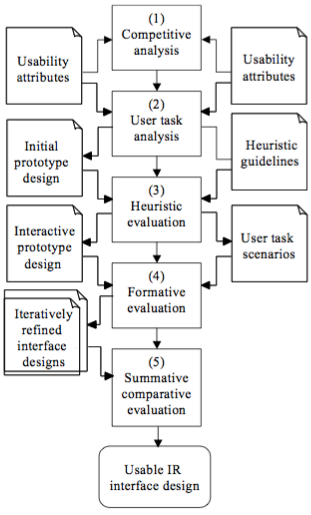
- Approccio user-centered di Ahmed
  - Analisi di uno o più sistemi di IR esistenti per effettuare test d'usabilità
  - Analisi dei task degli utenti durante il test d'usabilità
  - Progettazione delle prime soluzioni progettuali partendo dai risultati dell'analisi dei task
  - Valutazione euristica dei primi prototipi
  - Progettazione di prototipi interattivi, considerando gli spunti della valutazione euristica
  - Valutazione formativa dei prototipi interattivi usando scenari dei task
  - Rivisitare i prototipi progettati considerando la valutazione formativa
  - Valutazione sommativa dei prototipi finali e comparazione dei risultati con i risultati dell'analisi della concorrenza eseguendo task uguali

21

Progettazione e valutazione di interfacce per sistemi di IR

UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO

- Approccio user-centered di Ahmed



22

Un framework per la valutazione di sistemi di IR (1/3)

- Participants
  - HCI Experts (as usual)
  - Users
    - To investigate people information seeking needs
- Tasks
  - Formulation and submission of a query
  - Examination of the results
  - Possible feedback loop to re-formulate the query
  - Integration of search results and evaluation of the whole search

23

Un framework per la valutazione di sistemi di IR (2/3)

- Usage of realistic scenarios
- Simulated work task situation
- Usability measures
  - Effectiveness
    - interactive recall
    - interactive precision
    - interactive TREC precision
    - informativeness
    - cost
    - utility
  - Efficiency
    - the overall time the user takes
    - the time the user takes doing specific things
    - the time the user takes in specific or different modes
  - Satisfaction

24

Un framework per la valutazione di sistemi di IR (2/3)

UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO

Some Metrics from ISO 9241

Usability objective	Effectiveness measures	Efficiency measures	Satisfaction measures
Suitability for the task	Percentage of goals achieved	Time to complete a task	Rating scale for satisfaction
Appropriate for trained users	Number of power features used an expert user	Relative efficiency compared with power features	Rating scale for expert satisfaction
Learnability	Percentage of functions learned	Time to learn criterion	Rating scale for ease of learning
Error tolerance	Percentage of errors corrected successfully	Time spent on correcting errors	Rating scale for error handling

25

Un framework per la valutazione di sistemi di IR (3/3)

UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO

- Interaction measures
  - Number of queries
  - Number of search results viewed
  - Number of documents viewed
  - Number of documents saved
  - Query length
  - Appropriate combinations of the above measures
- User characteristic measures
  - Sex, age, profession, computer experience, search
  - Experience, Internet perceptions, cognitive style, etc.
  - Preference
  - Mental effort and cognitive load
  - Flow and engagement
  - Learning and cognitive transformation

26

Information Visualization nei sistemi IR

- La visualizzazione dei risultati dei sistemi IR è un settore molto studiato dall'InfoVis
- I risultati delle query tradizionalmente non includono
  - Quanto la query è correlata ai documenti
  - La frequenza dei termini
  - Come i termini cercati sono distribuiti nei documenti trovati
  - Lunghezza dei documenti
- Il ranking dei documenti è nascosto
- Impossibilità di comparare più risultati
- Impossibilità di effettuare ricerche più complesse

27

Information Visualization nei sistemi IR

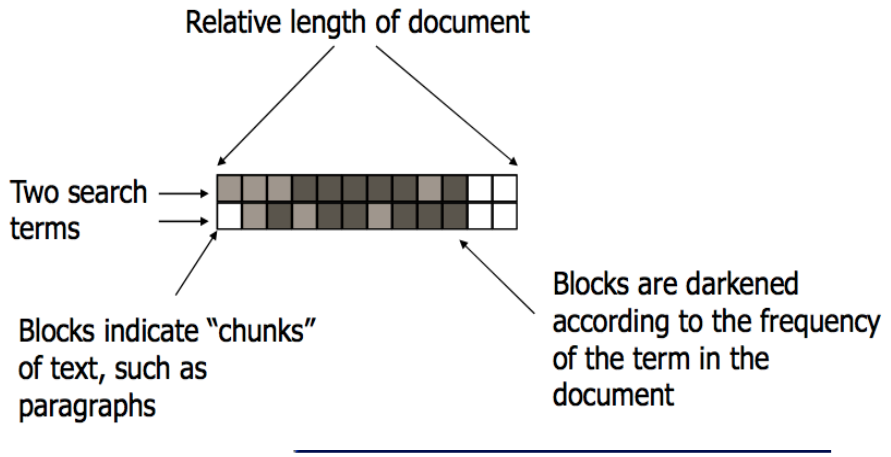
TileBars

- Obiettivo
  - Minimizzare il tempo e lo sforzo per decidere quali documenti si vuole vedere in dettaglio
- Idea
  - Mostrare il ruolo dei termini della query nei documenti ritrovati
- Rappresentazione grafica della disposizione e sovrapposizione dei termini
- Contemporaneamente indica:
  - Lunghezza relativa del documento
  - Frequenza dei termini nel documento
  - Distribuzione dei termini rispetto ai documenti e agli altri termini

28

Information Visualization nei sistemi IR

TileBars



29

Information Visualization nei sistemi IR

TileBars: problemi e limiti

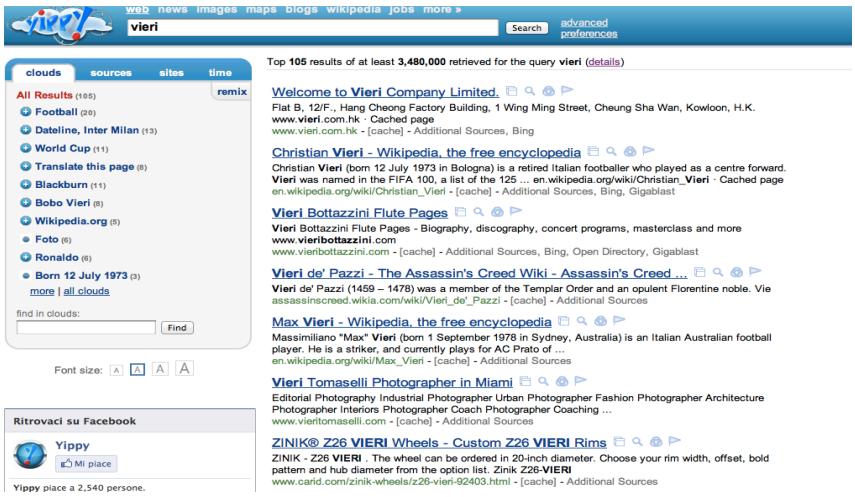
- La visualizzazione orizzontale non rispecchia bene il modello mentale
- La visualizzazione non si adatta bene a ricerche effettuate con molte parole

30

Information Visualization nei sistemi IR

UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO

Web search clustering



31

Information Visualization nei sistemi IR

UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO

Sparkler

- Visualizzazione più astratta dei risultati
- Mostra la 'distanza' tra la query e i documenti per far capire meglio la vicinanza tra essi
- Mostra anche i risultati dei documenti in risposta a piú query contemporaneamente

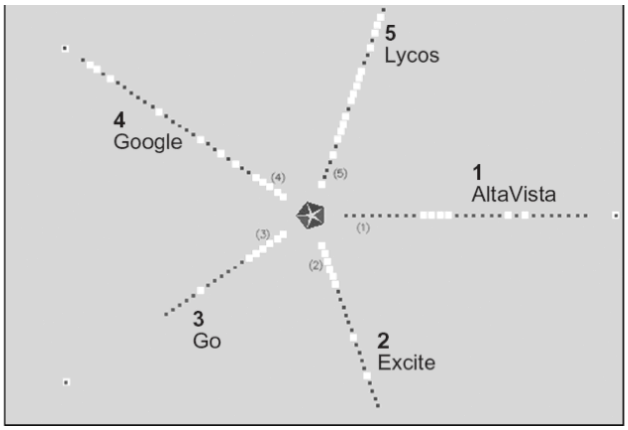
La visualizzazione in sparkler

- Triangoli – query
- Quadrati – documenti
- La distanza tra triangoli e quadrati rappresenta la rilevanza dei documenti rispetto alla query

32

Information Visualization nei sistemi IR

Sparkler:



Use it to compare search results from different search engines

33

Information Visualization nei sistemi IR: oltre la semplice ricerca

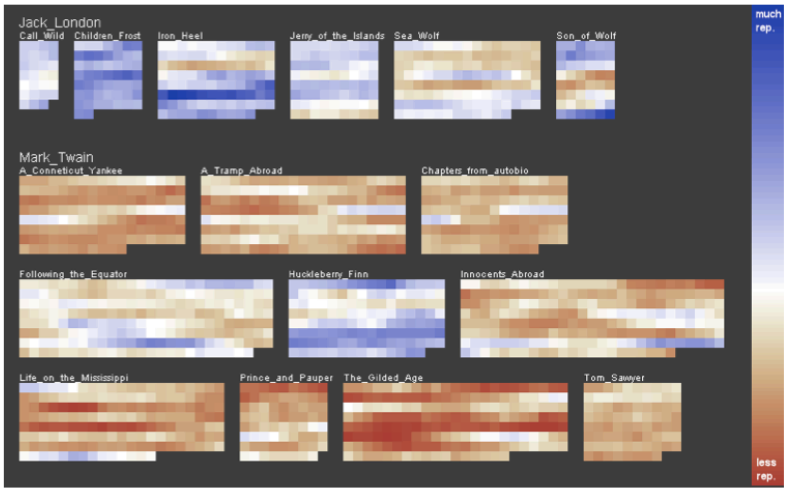
- Oltre i semplici risultati della ricerca, si vuole
  - Presentare più contesto dei documenti
  - Presentare più informazioni dei documenti
  - Sintetizzare il contenuto dei documenti

34

Information Visualization nei sistemi IR: oltre la semplice ricerca

UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO

- L'impronta digitale dei documenti



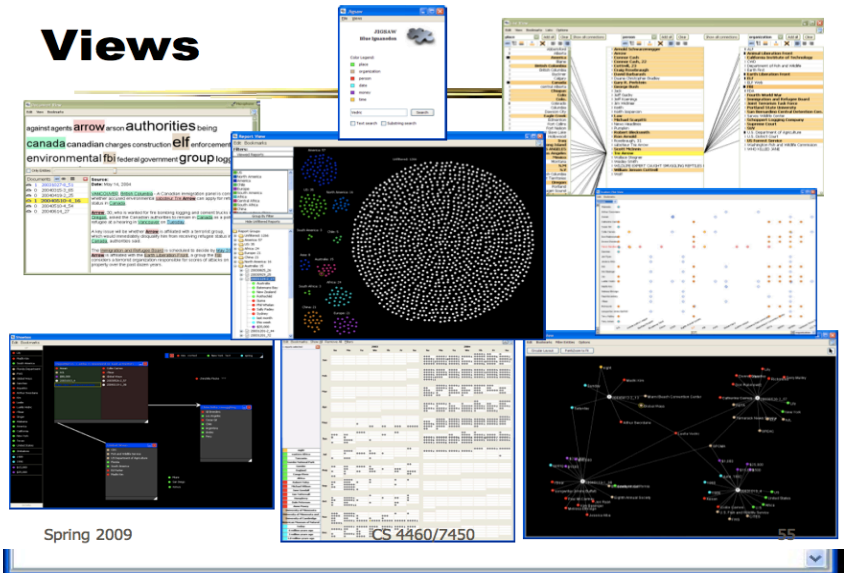
(d) Simpson's Index

35

Information Visualization nei sistemi IR: oltre la semplice ricerca

UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO

- **Views**



36

Information Visualization nei sistemi IR: visualizzazione di concetti e temi

- Un altro obiettivo è la comprensione di concetti e temi in collezioni di documenti
- Il problema e la sfida è come presentare i contesti/semantica/ temi dei documenti a qualcuno che non ha tempo per leggerli
- A chi potrebbe interessare?
  - Ricercatori, giornalisti, CIA, InfoVis, mondo giuridico, etc.

37

Information Visualization nei sistemi IR: visualizzazione di concetti e temi

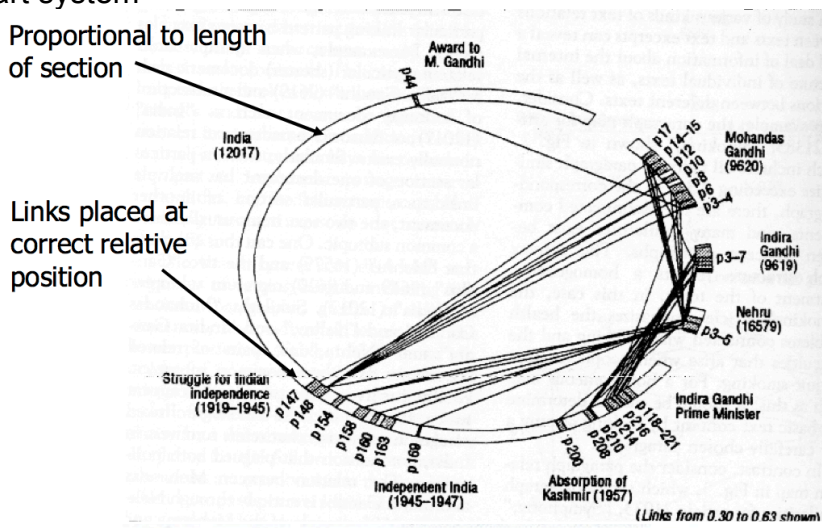
Smart system

- Usa il vector space model per i documenti
- Divide i documenti in capitoli e sezioni e le tratta separatamente come atomi
- Plotta atomi dei documenti sulla circonferenza dei cerchi
- Disegna linee tra oggetti se la loro similarità è superiore a una determinata soglia

38

Information Visualization nei sistemi IR: visualizzazione di concetti e temi

Smart system



39

Information Visualization nei sistemi IR: visualizzazione di concetti e temi

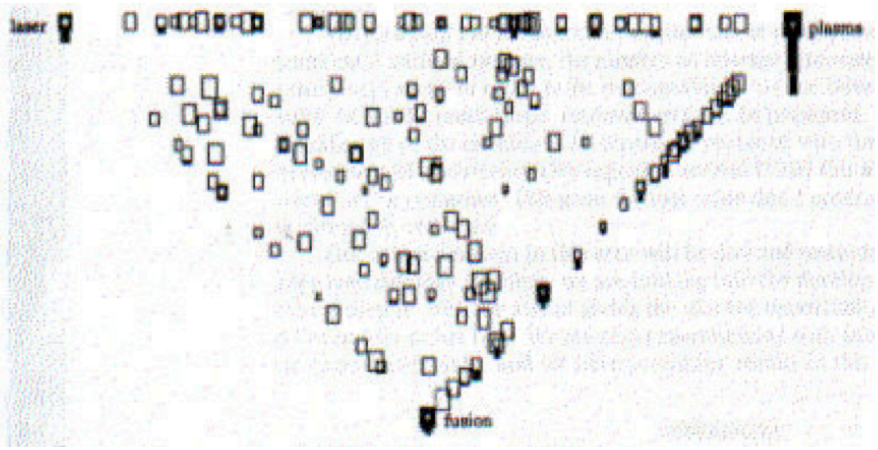
Relazioni tra documenti e loro argomenti

- L'Idea è di capire il contenuto dei documenti e capire come sono collegati l'uno l'altro
- L'utente deve inserire delle keyword d'interesse
- Gli argomenti (keyword) inseriti dagli utenti sono i vertici del triangolo
- I documenti sono i punti all'interno del triangolo
- La distanza dai vertici indica quanto un documento è correlato all'argomento del vertice

40

Information Visualization nei sistemi IR: visualizzazione di concetti e temi

Relazioni tra documenti e loro argomenti



41

Information Visualization nei sistemi IR: visualizzazione di concetti e temi

Relazioni tra documenti e loro argomenti pro e contro

Pro

- Comunica bene la relazione tra documenti e argomenti
- Metodologia semplice da comprendere
- Può plottare anche insiemi medio/grandi di documenti

Contro

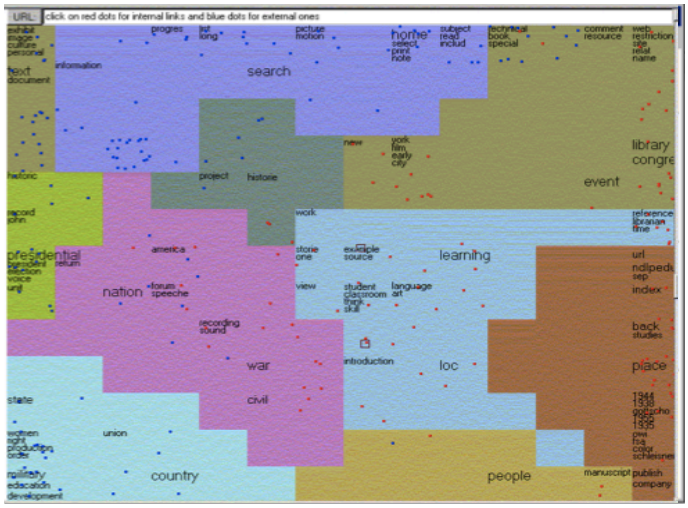
- Non dice nulla riguardo i singoli documenti
- I singoli documenti perdono i dettagli
- Diventa troppo complesso e incomprensibile con collezioni di documenti grandi

42

Information Visualization nei sistemi IR: visualizzazione di concetti e temi

UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO

Mappe dei concetti e dei documenti

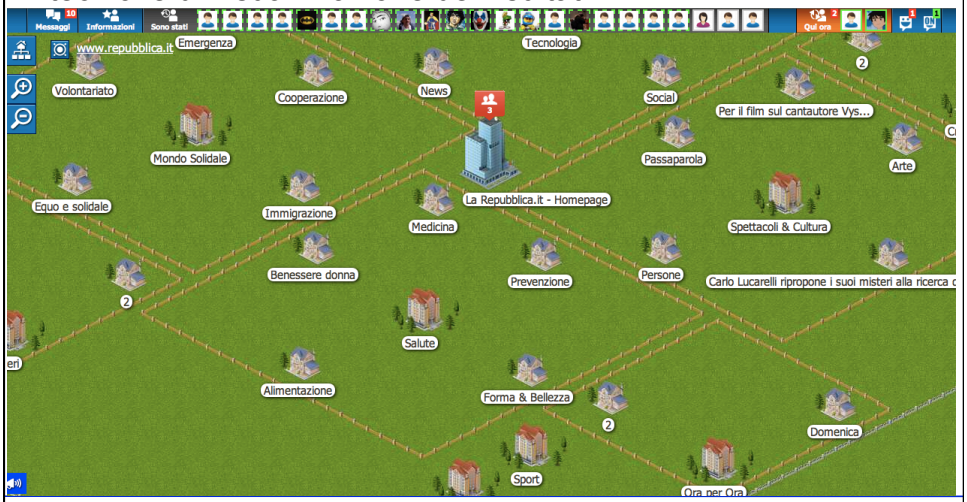


43

Information Visualization nei sistemi IR

UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO

- Volunia: un nuovo motore di ricerca tutto italiano con nuove tecniche di visualizzazione dei risultati



44

Information Visualization nei sistemi IR

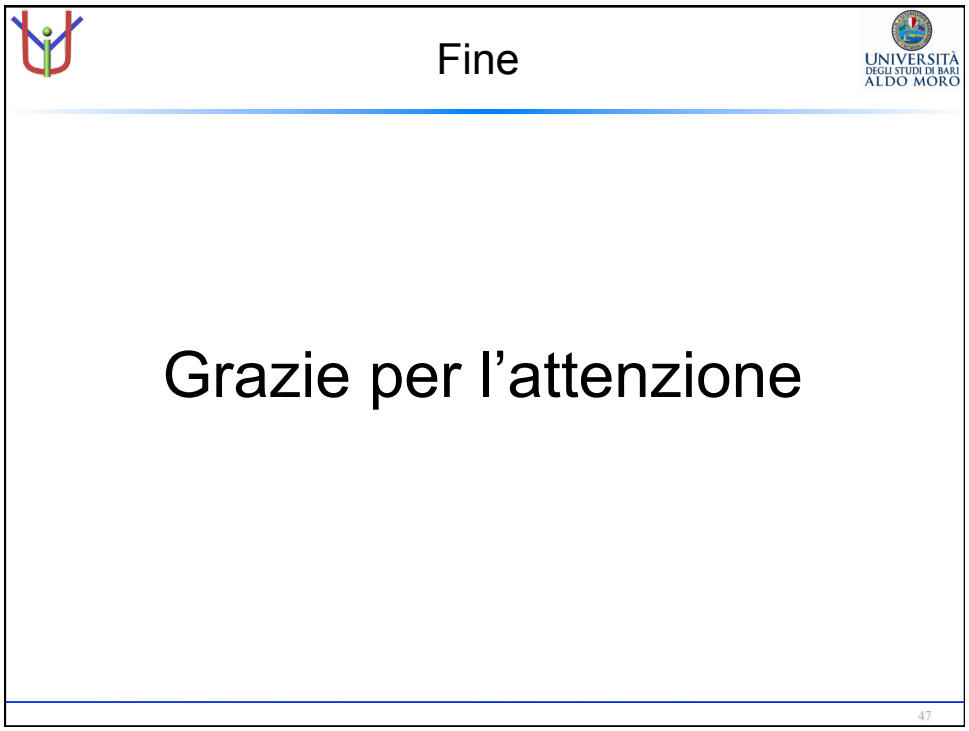
- Many eyes: un sistema per visualizzare le informazioni ricercate in molteplici modi
- http://www-958.ibm.com/software/data/cognos/manyeyes/page/Visualization_Options.html (Tipi di visualizzazioni)
- http://www-958.ibm.com/software/data/cognos/manyeyes/page/create_visualization.html (Creare le visualizzazioni)

45

Riferimenti

- 1)Basi dell'IR
  http://www.diit.unict.it/users/alongheu/sei2/aa0910/sei2_lezione10_information_retrieval.pdf
- 2) *HCI View of Information Retrieval Evaluation*
  http://www.promise-noe.eu/documents/10156/ff7f0168-088b-4c52-90c6-bd405fca345e
- 3)Dettaglio di un framework per la progettazione user-centered di interfacce IR
  http://ir.inflibnet.ac.in/dxml/bitstream/handle/1944/1347/2.pdf?sequence=1
- 4)Articolo di JigSaw del Prof. Stasko
  http://www.cc.gatech.edu/~john.stasko/papers/iv08-jigsaw.pdf

46


Fine
UNIVERSITÀ
DEGLI STUDI DI BARI
ALDO MORO
Grazie per l'attenzione
47