Sistemi per il recupero delle informazioni

Gabriele Pozzani

A.A. 2014/2015

Corso di Laurea Magistrale in
Editoria e Giornalismo

Interfacce utente per il recupero delle informazioni

L'interfaccia utente

- È la finestra tramite cui si interagisce con un SRI
- Ha l'obiettivo di aiutare nel comprendere i bisogni degli utenti
- Inoltre dovrebbe aiutare gli utenti nel
  - Formulare le loro interrogazioni
  - Selezionare tra le sorgenti di informazioni disponibili
  - Comprendere i risultati delle ricerche
  - Tenere traccia dei progressi nelle ricerche

Come gli utenti cercano (I)

- L'interazione tra gli utenti e i SRI dipende da
  - Il tipo di richiesta
  - L'esperienza e la conoscenza dell'utente
  - La quantità di tempo a disposizione
- Vi sono due tipi principali di ricerche
  - Information lookup
  - Exploratory search

Information lookup

- La ricerca di un particolare fatto o risposta
- Soddisfatta da frammenti ben definiti di informazione: numeri, date, nomi, siti web
- Funziona bene con i classici motori di ricerca

Exploratory search

- Si divide ulteriormente in
  - Learning search
    - Richiede più di un'interrogazione e di un risultato
    - Richiede tempo
      - Per scorrere e leggere diversi risultati
      - Per trasformare quanto letto in nuova conoscenza
  - Investigating: processo di lungo periodo
    - Richiede più sessioni di ricerca su un lungo periodo di tempo
    - Può richiedere il recupero di una grossa parte dei documenti rilevanti disponibili

Come gli utenti cercano (II)

- La ricerca di informazioni può essere vista come parte di un processo più generale di *sensemaking*
  - Processo iterativo che porta a formarsi una rappresentazione concettuale di una grande collezione di informazioni

Ricerca classica vs dinamica

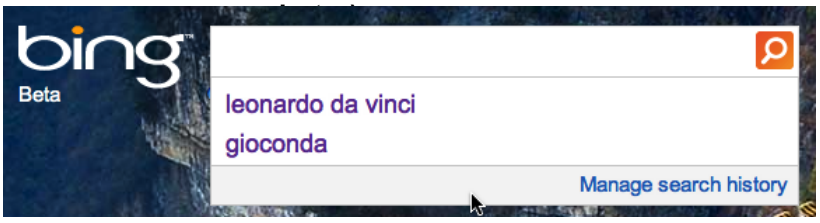
- Nella definizione classica di ricerca dell'informazione si ha il seguente processo
  - Identificazione del problema
  - Strutturazione del bisogno informativo
  - Formulazione della query
  - Analisi dei risultati
- Recentemente si è enfatizzata la natura dinamica delle ricerche secondo cui
  - Gli utenti imparano mentre cercano
  - Il bisogno informativo evolve durante la ricerca e l'analisi dei risultati intermedi

Orienteering

- L'attuale strategia di ricerca degli utenti web è detta orienteering ed è parte dell'approccio dinamico
  - Eseguire una prima ricerca
  - Osservare i risultati ottenuti
  - Raffinare o riformulare la query sulla base dei risultati precedenti
- Il 52% degli utenti modifica la propria interrogazione

Il processo di ricerca (I)

- Numerosi studi sono stati effettuati su come le persone affrontano il problema della ricerca di informazioni
- Risultati comuni in queste ricerche sono
  - Gli utenti spesso riformulano le query apportando piccole modifiche
  - Spesso gli utenti cercano qualcosa che hanno già cercato in passato (il 33% delle query sono ripetute [Yahoo!])
- Per questo diversi SRI includono nell'interfaccia
  - la storia delle ricerche passate (e.g., Bing)
  - la possibilità di riformulare velocemente la query (ricerche



Il processo di ricerca (II)

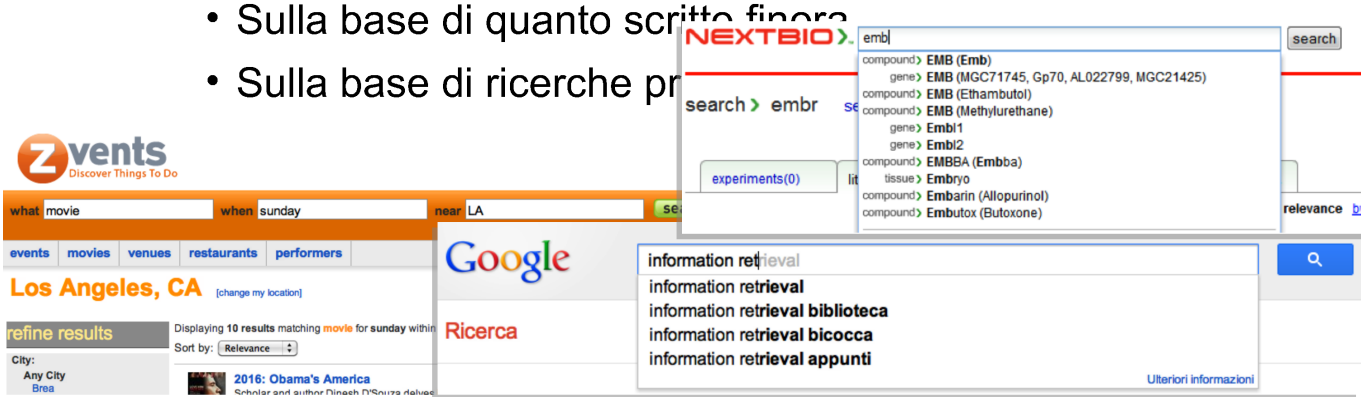
- Inoltre si è osservato che gli utenti
  - Valutano con difficoltà se un documento è rilevante o no rispetto ad un argomento
    - La difficoltà aumenta con la minore conoscenza dell'argomento
  - Tendono ad osservare solo i primi risultati più rilevanti
  - Tendono a pensare che il primo o i primi due risultati siano migliori di quelli successivi
  - Valutano con difficoltà quanta porzione dei documenti rilevanti essi hanno trovato
  - Usano diverse strategie a seconda dell'esperienza
    - Gli utenti esperti sono più pazienti
    - Questo atteggiamento positivo porta ad ottenere migliori risultati nelle ricerche

Interfaccia

- La classica interfaccia di un SRI include un campo testuale
  - La dimensione del campo è correlata alla lunghezza delle query
    - Un campo piccolo scoraggia query lunghe
    - Un campo lungo incoraggia query lunghe

Funzionalità

- Il campo di ricerca può includere diverse funzionalità
  - Aggiunta di un ulteriore campo per filtrare subito la ricerca
  - La capacità di interpretare richieste temporali come “domani”, “mercoledì” (zvents)
  - Auto-completamento, suggerimenti
    - Sulla base di quanto scritto finora
    - Sulla base di ricerche pr

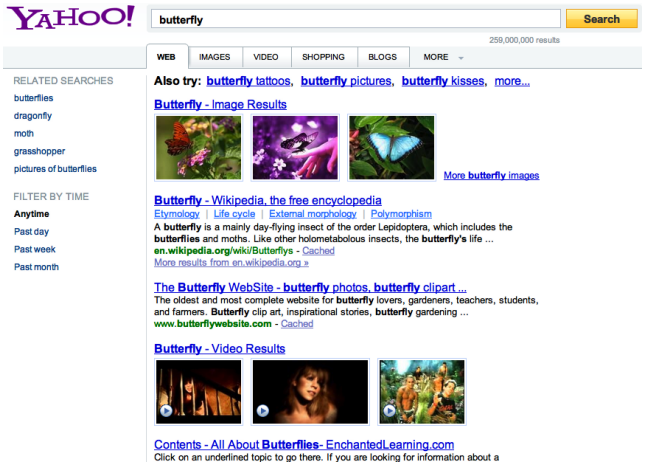


Visualizzazione dei risultati (I)

- I documenti recuperati possono essere mostrati
  - Completamente
  - Un qualche surrogato
- Il surrogato mostrato all'utente “riassume” il documento
  - È una scelta fondamentale nel successo di un SRI
  - La sua scelta e qualità influenza la rilevanza percepita dall'utente

Visualizzazione dei risultati (II)

- Tipici surrogati sono
  - Il titolo del documento
  - Metadati (data di pubblicazione, autore, URL)
  - Un testo estratto dal documento stesso (snippet)
    - Dove appaiono i termini ricercati (keywords in context)
- Diversi tipi di risultati possono essere mostrati contemporaneamente
  - Documenti testuali
  - Video
  - Immagini



Visualizzazione dei risultati (III)

- Quando il surrogato contiene l'informazione cercata si parla di “answer engine”
- Determinare quale testo mostrare nel sommario di un risultato è un problema centrale
  - Mostrare brevi estratti che contengono i termini cercati
  - Mostrare frasi complete e contigue
    - Aiuta a comprendere il contesto e valutare meglio la rilevanza del risultato
    - Frasi troppo lunghe però sono poco desiderabili
    - Trade-off

Riformulazione delle query (I)

- Come detto, l'interazione con il SRI avviene solitamente tramite una riformulazione della query sulla base dei risultati ottenuti nella query precedente
  - Alcuni SRI forniscono apposite funzionalità a tale fine
    - Correzione degli errori di battitura
    - Suggerimenti
    - Ricerche correlate

Riformulazione delle query (II)

- Suggerimenti e ricerche correlate sono ampiamente adottate dai SRI e usate
  - Circa il 6% degli utenti clicca sulle ricerche suggerite
  - Circa l'8% delle query è generata dai suggerimenti
- Suggerimenti e ricerche correlate possono essere basate:
  - Sulle ricerche passate dello stesso utente
  - Sul comportamento di altri utenti
    - Query simili effettuate da altri utenti
    - Usare termini estratti dai documenti cliccati da altri utenti che hanno effettuato la stessa ricerca

Riformulazione delle query (III)

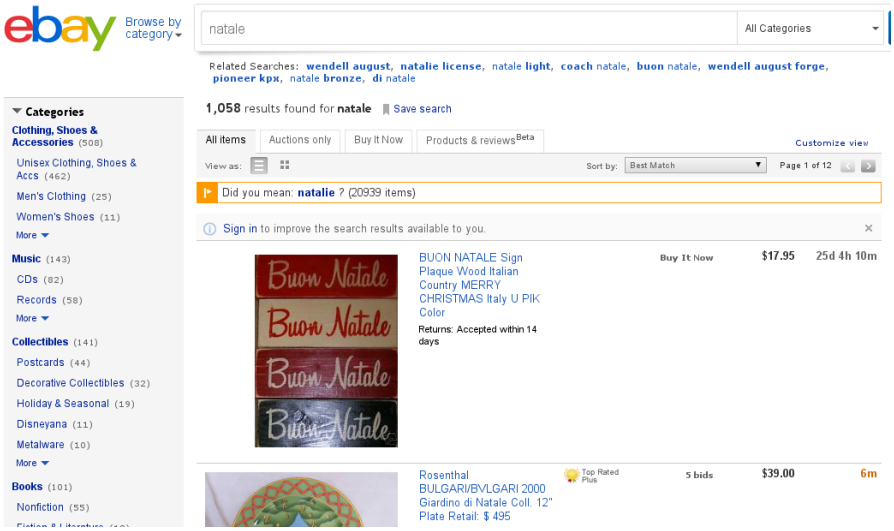
- Un'altra tecnica è quella di richiedere all'utente un feedback
  - Permettere all'utente di indicare quali documenti recuperati sono rilevanti
  - Il SRI calcola una nuova query sulla base del feedback
- Poco o per nulla usata
  - Gli utenti non sono bravi nel valutare la rilevanza dei documenti recuperati

Organizzare i risultati

- Per gli utenti può essere utile ottenere i risultati organizzati in gruppi
  - Categorie
  - Cluster
- Categorie
  - “etichette” che rappresentano concetti rilevanti
  - Meglio se “complete” (coprono il maggior numero di casi)

Categorie (I)

- Vi sono tre tipi di suddivisione in categorie:
  1)Flat: semplice lista di argomenti/soggetti
  - Possono essere usate sia per raggruppare che per filtrare i risultati



Categorie (II)

- Vi sono tre tipi di suddivisione in categorie:
  2)Gerarchica: risultati suddivisi in categorie e queste in sottocategorie
  - Utilizzata da alcuni dei primi motori di ricerca
  - Non più utilizzata nei motori di ricerca web
  - Utilizzata in alcuni SRI per libri
    - La gerarchia è data dalla suddivisione in capitoli, sezioni, … dei libri

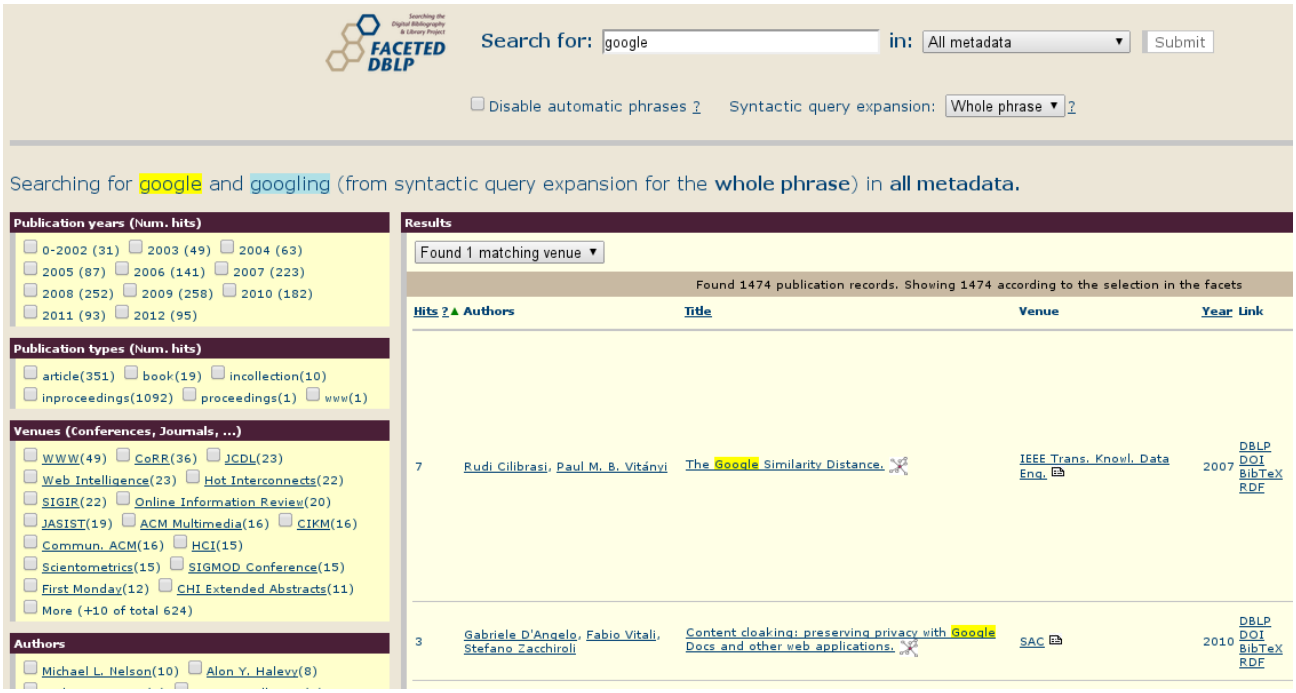
Categorie (III)

- Vi sono tre tipi di suddivisione in categorie:

  3)Faceted (dimensionale):

  - Al contrario delle categorie flat, le faceted permettono ad uno stesso risultato di appartenere a più categorie
  - Ogni categoria corrisponde ad una dimensione/caratteristica dei risultati

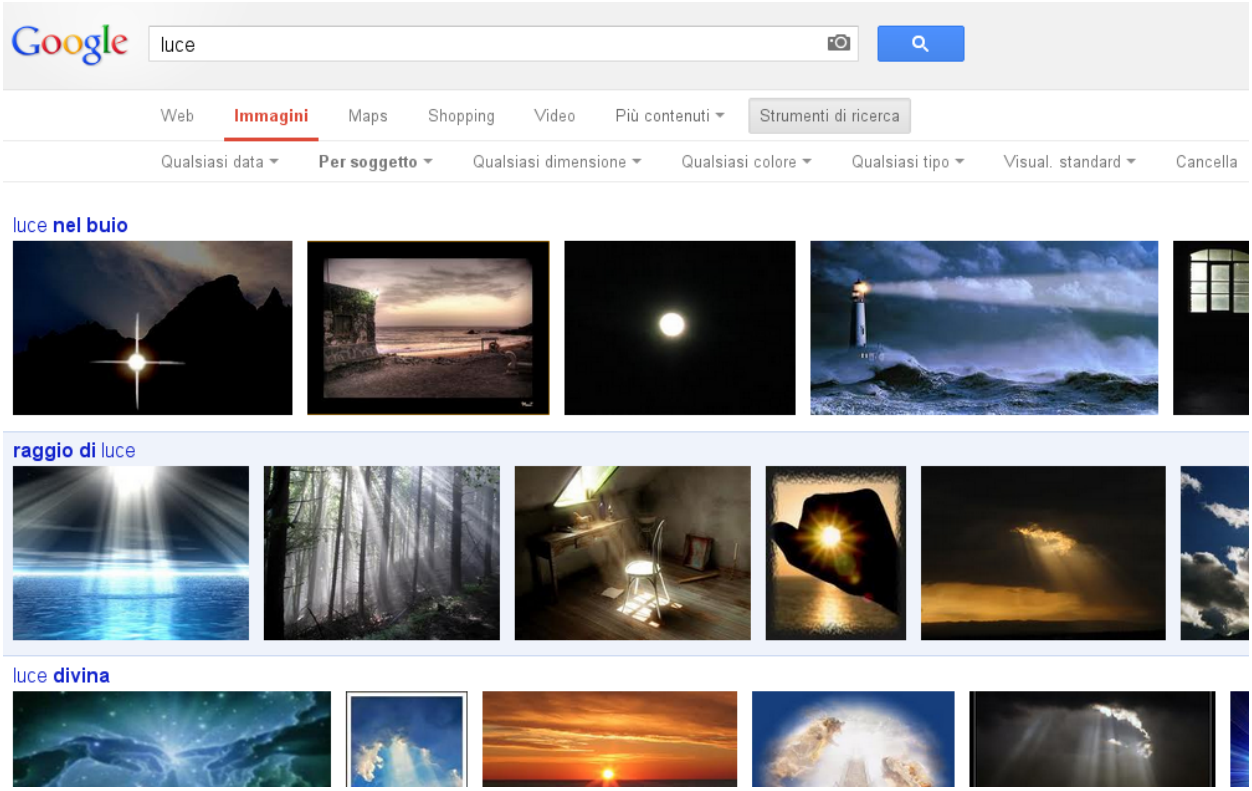
Faceted search, esempio



Clustering

- Raggruppamento dei risultati in gruppi rispetto ad un qualche tipo di “similarità”
  - I gruppi (cluster) sono decisi dinamicamente in base alla ricerca e ai risultati
  - I gruppi sono decisi in modo completamente automatico

Clustering, esempio



La progettazione

- La progettazione dell'interfaccia utente rientra nel campo dell'Interazione Uomo-Macchina (HCI)
  - Basato sullo studio di cosa le persone pensano della tecnologia, come la utilizzano e come vi rispondono
  - È uno studio incentrato sugli utenti
- È un processo iterativo
  - L'interfaccia è solitamente continuamente modificata per venire in contro a nuove esigenze, capacità, funzionalità

Valutazione (I)

- Anche la valutazione di un'interfaccia ruota attorno agli utenti
  - Non esistono misure numeriche precise per valutarla in modo oggettivo
  - La qualità di un'interfaccia dipende da come gli utenti la percepiscono e vi rispondono
- Le ragioni in base alle quali gli utenti valutano un'interfaccia sono varie
  - Velocità, FAMILIARITÀ, estetica, funzionalità preferite, accuratezza percepita dei risultati

Valutazione (II)

- La valutazione avviene solitamente tramite esperimenti
  - Studio longitudinale
    - I partecipanti all'esperimento utilizzano una nuova interfaccia per un “lungo” periodo di tempo
    - Il loro comportamento/utilizzo viene monitorato e registrato
    - La nuova interfaccia viene valutata basandosi sui dati registrati, su questionari e interviste agli utenti

Valutazione (III)

  - A/B testing: utilizzato per sistemi già ampiamente in uso
    - Ad un gruppo di utenti casuali viene fatta utilizzare una nuova interfaccia
    - Il loro comportamento viene monitorato e registrato
    - I dati registrati sulla nuova interfaccia vengono confrontati con dati provenienti ad un altro gruppo di utenti casuali che ha continuato ad usare l'interfaccia vecchia