Intelligent Web

1 – introduction & definitions

Fabio Gasparetti
gaspare@dia.uniroma3.it

Overview

- Motivations & Goals
- Definitions
  - Information
  - Information Overload
  - Information Needs
  - Information Seeking
  - Information Filtering
  - Information Retrieval

Motivations & Goals

- La *HCI* (Human-Computer Interaction) con grosse collezioni di informazioni pone il problema della ricerca degli elementi informativi che soddisfino i *bisogni informativi* attuali dell'utente
- *Internet* rende accedibili grossi moli di dati a cui ognuno può facilmente accedere

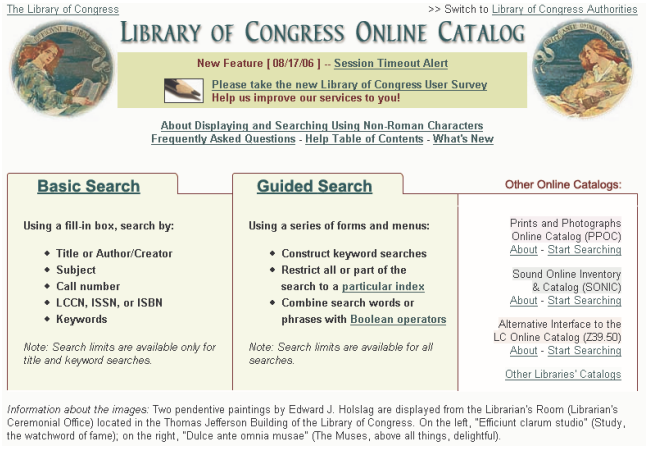
Online Information Services

www.loc.gov

Oldest US cultural institution and largest library in the world:

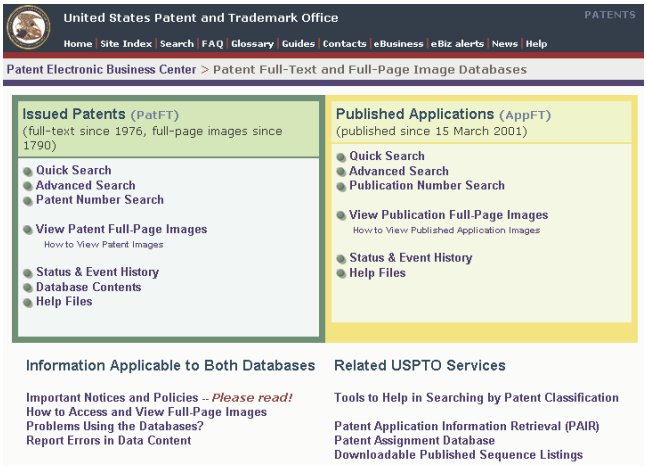
·130M items
·853 km of bookshelves
·29M books and printed materials
·2.7M recordings
·12M photographs
·4.8M maps,
·58M manuscripts.

…come se stessimo al centro di una libreria a cerchio di raggio 135km

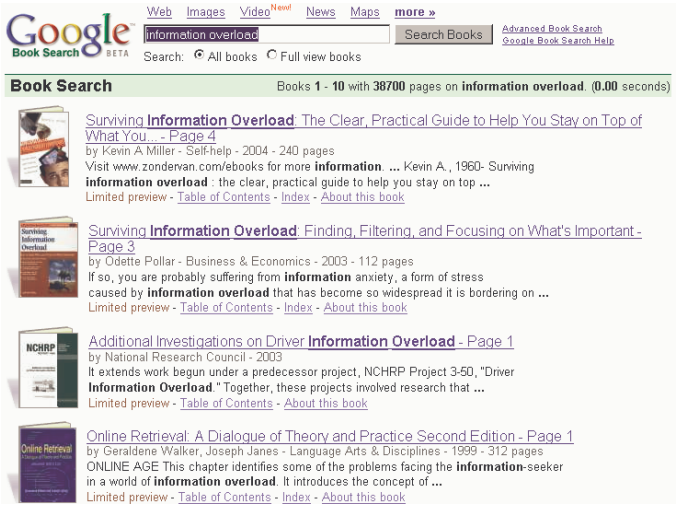


Online Information Services

www.uspto.gov



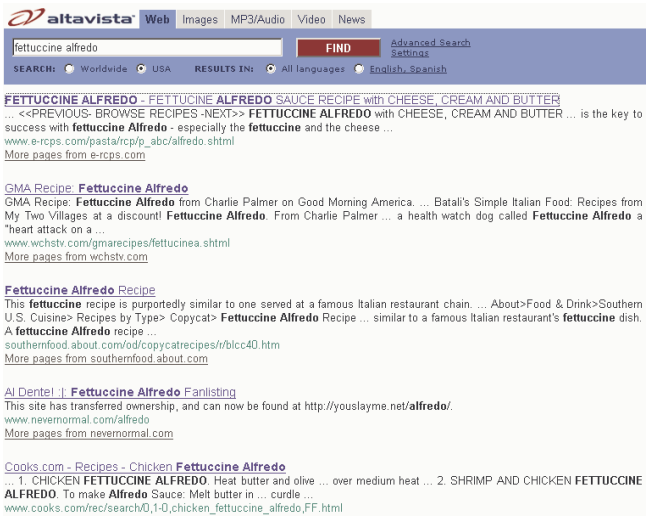
Online Information Services



Online Information Services

Search Engines:

•11.5B indexable pages



Online Information Services

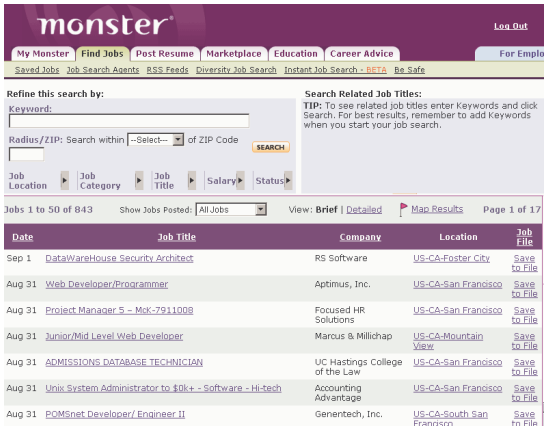
- …e la dimensione del Web?
- Se i motori di ricerca riescono a indicizzare $11 \cdot 10^9$ di pagine ([a]), allora la dimensione del Web e':
  1. ~ $11 \cdot 10^9$ pagine?
  2. ~ $50 \cdot 11 \cdot 10^9$ pagine?
  3. ~ $550 \cdot 11 \cdot 10^9$ pagine?
- *Deep Web* = Web ([3]) – Indexable Web ([a])

Online Information Services

Deep Web: pages under search boxes (1/2)

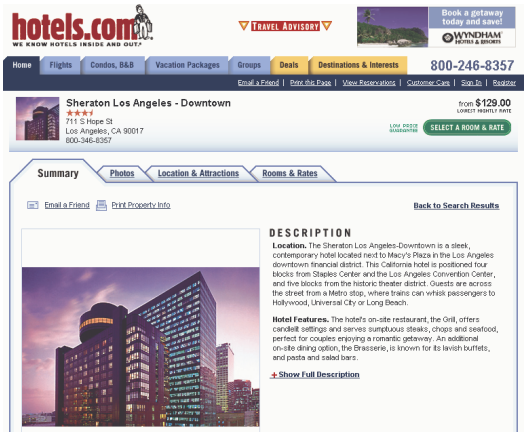
Online Information Services

Deep Web: pages under search boxes (2/2)



Online Information Services

Deep Web: robot excluded pages



Online Information Services

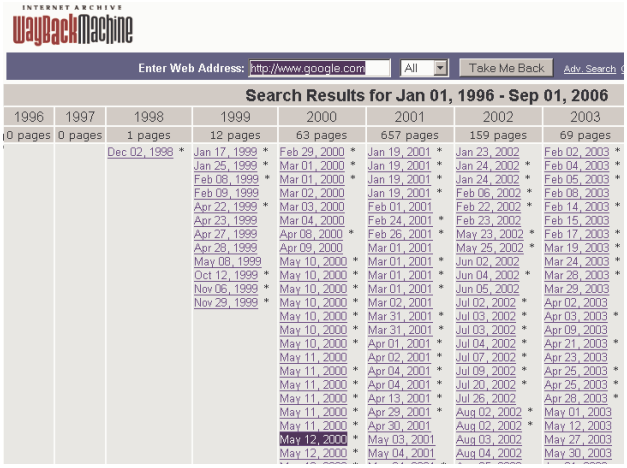
Deep Web: no textual data

Online Information Services

www.archive.org

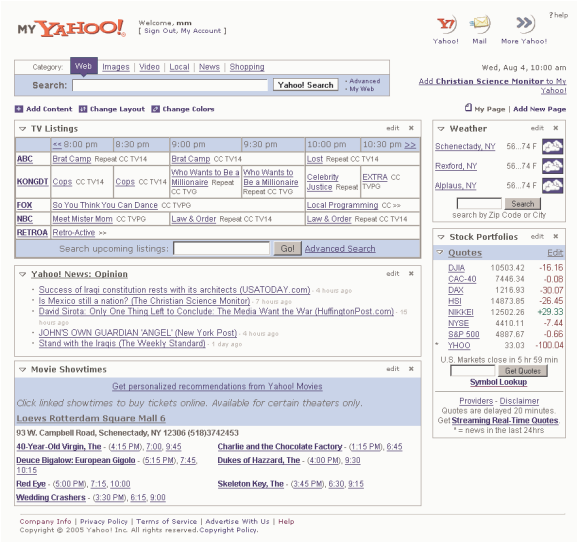
Internet Archive: purpose of offering permanent access for researchers, historians, and scholars to historical collections that exist in digital format and to prevent the Internet and other "born-digital" materials from disappearing into the past

- **55B Web pages**
- **30K text items**
- **38K audio items**
- **40K moving image items**
- **34K software items**



Online Information Services

Online Information Services



Motivations & Goals

- Obiettivo: ridurre il *tempo* e le risorse (anche cognitive) impiegate dall'utente per la ricerca aumentando la probabilità di successo, filtrando le informazioni più attinenti in base agli interessi dell'utente
- Approccio: personalizzare l'interazione tra utente e sistema informativo
- Domini principali: grosse librerie digitali e il Web

Overview

- Motivations & Goals
- Definitions
  - Information
  - Information Overload
  - Information Needs
  - Information Seeking
  - Information Filtering
  - Information Retrieval

Information

- "Any difference you perceive, in your environment or within yourself. It is any aspect that you notice in the pattern of reality"

→ If the world is entirely known, there is no information

Information

- Information may be used to indicate either
  - a process (informing),
  - what is perceived (the knowledge communicated),
  - objects that are informative (text,audio)
- Other definitions:
  - *specific requirements*, e.g., true, useful
  - *top downs*: abstract different definitions, e.g., a message expressed in some medium, it has the potential of altering a person's consciousness
  - *bottom ups*: list exhaustively all the possible forms, e.g., book, Web page, radio broadcast, conversation

Information vs Data

- "a datum is factual[1] information (as measurements or statistics) used as a basis for reasoning, discussion, or calculation"
- a datum is "information in numerical form that can be digitally transmitted or processed"

(1) concerned with what is actually the case rather than interpretations of or reactions to it

Information vs Datum

- 4 9 0 1 0 7 1 2 3 4 5 6 7 8 9 9 ?
  - Germany phone number?
  - Lock code?
  - Milliseconds after 1970?

Information vs Knowledge

- "knowledge is information that has been *sifted*, *organized* and *understood* by a human brain"
- "information is acquired by being told, whereas knowledge can be acquired by thinking" (Machlup '83)
- "knowledge is internal; it cannot be received but must be internally created" (Hayes '93)
- → knowledge is obtained from investigation, study, or reasoning of data/information

Overview

- Motivations & Goals
- Definitions
  - Information
  - Information Overload
  - Information Needs
  - Information Seeking
  - Information Filtering
  - Information Retrieval

Information Overload

- "La montagna di pubblicazioni scientifiche è in costante *aumento*. Ma è tangibile la preoccupazione di impantanarsi nelle ricerche a causa del crescente numero di possibili specializzazioni […] *la difficoltà non è nella eccessiva mole di pubblicazioni* […] *ma piuttosto nella nostra abilità di gestirle proficuamente*" (Vannevar Bush "As We May Think" '45)
- E.g.,on average 2/3 of the newspaper stories are ignored by the public (Graber '84)

Information Overload

- "is the state of an individual or system in which excessive communication inputs cannot be processed, leading to breakdown (effects → failure to function)" (Rogers '86)
- An event that usually occurs during the "filtering" of the environmental inputs (when)

Information Overload

- *"What information consumes is rather obvious: it consumes the attention of its recipients.*
  Hence a wealth of information creates a poverty of attention and a need to allocate that attention efficiently among the overabundance of information sources that might consume it" (*Herbert A. Simon* 1978)

Information Overload

- Several user response categories to overload (Miller et.al'60):

dysfunctional:

1. *Omission*: failing to process some of the inputs
2. *Error*: processing the information incorrectly in some way
3. *Escaping*: giving up the burden of attending the inputs entirely

can be dysfunctional if not based on well-considered priorities:

4. *Queuing*: delaying the processing of some information with the intention of catching up later
5. *Approximation*: lowering standards of discrimination by being less precise in categorizing inputs and responses
6. *Multiple Channels*: splitting up the incoming information in order to decentralize the response

Overview

- Motivations & Goals
- Definitions
  - Information
  - Information Overload
  - Information Needs
  - Information Seeking
  - Information Filtering
  - Information Retrieval

Information Needs

- "it's a recognition that your knowledge is *inadequate* to satisfy a goal that you have" (goal-centered)
- *ASK:* anomalous state of knowledge: "an anomaly (gap/uncertainty) in the state of knowledge regarding a situation or topic"
  (Belkin & Croft '87)
- "inner motivational state (e.g., wanting, believing, doubting, fearing, expecting) that causes thought and action" (emotive) (Grunig '89)

→ Sono legati alla conoscenza/background interni dell'utente

Information Needs

- Some characteristics of the *needs*:
  1. instrumental: it involves reaching a desired goal
  2. I maybe unaware of it: if i don't know exactly where or what to search

Information Needs

- Why people come to ask questions at library desk? We can recognize 4 stages: (Taylor '68)

formalization & comprehension ↓

visceral ← *abstract/vague*
conscious
formalized
compromised ← *tangible/definite*

Information Needs

1. visceral: "conscious or even unconsciuous need for information… a vague sort of dissatisfaction…probably inexpressible in in linguistic terms"
2. conscious: "a conscious mental description…an ambiguous and rambling statement" which sometimes results in talking to another person about it
3. formalized: qualified and rational statement of the need; the person is not aware whether the need could be answered in that form by any available person or information system
4. compromised level: where the need may be a question asked of a librarian or a search system; the question reflects the kinds and forms of data that may be available (books,images,etc).

Information Needs

- Riconoscere gli user needs è semplice se ci troviamo nel *formalized/compromised level*, anche se: “people are very bad at introspecting and then reporting on their cognitive processes” (Nisbett and Wilson ‘79)
- Nei livelli *conscious* o *formalized* è possibile pensare di interagire con l’utente interpretando l’input (la query è una rappresentazione imperfetta dei needs)
- In tutti i livelli è possibile osservare il comportamento dell’utente e tentare di capire i relativi needs

Information Needs

- In alcuni casi vengono suddivisi in categorie in base alla durata:
  - a breve termine: estemporanei; necessari per un certo task che ha una scadenza prefissata, e.g., meteo, conferme voli
  - a lungo termine: stabili o lentamente variabili nel tempo, che si possono eventualmente ripetere spesso; fanno di solito riferimento a task di lunga durata (come intraprendere una certa carriera), e.g., “java SDK”, “arte ‘700”

Overview

- Motivations & Goals
- Definitions
  - Information
  - Information Overload
  - Information Needs
  - Information Seeking
  - Information Filtering
  - Information Retrieval

Information Seeking

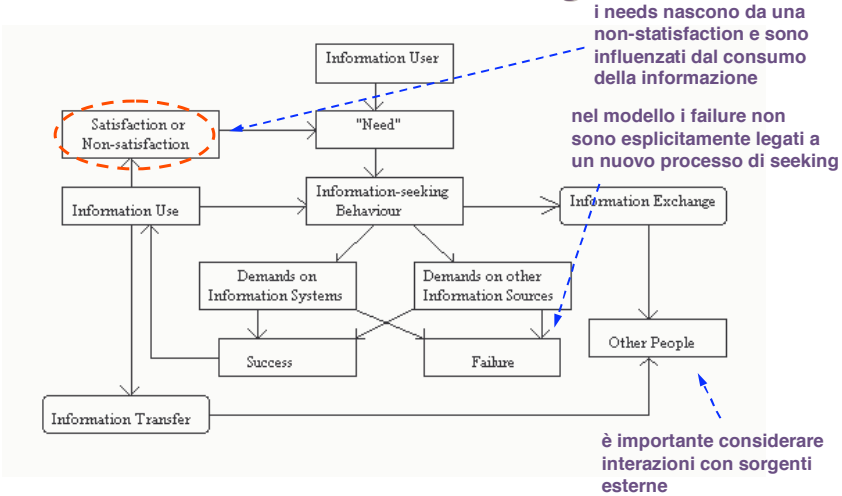
- "It takes place when a person recognize a *gap in their knowledge* that may motivate that person to acquire new information" (Zerbinos '90)
- "a process in which humans purposefully engage in order *to change their state of knowledge* closely related to learning and *problem solving*" (Marchionini '95)
- "*purposive* acquisition of information from selected information carriers (Johnson '97)

Information Seeking vs Behaviour

- Information Behaviour: “the totality of human behaviour in relation to sources and channels of information, including both active and passive information seeking, and information use […]

  it includes face-2-face communication with others, as well as the passive reception of information as in (e.g., watching tv ads) without any intention to act on the information given”
  (T.Wilson ’99)

Information Seeking



“Models in information behaviour search” di T.D.Wilson. J.Documentation ’99

Information Seeking

- Nel campo della ICT i *processi* di seeking (visti come “activity conducted by humans, perhaps with the assistance of a machine”) possono essere organizzati in base al tipo di bisogno/information source.
- Information sources possono contenere/generare diverse tipologie di dati: (hyper)testo, audio, immagini, video, etc.

Information Seeking

- *Need stabile/dinamico* temporalmente (anche rispetto alle nuove informazioni rese disponibili)
- *Need specifico/vario*
- *Source stabile/dinamica* temporalmente
- *Source strutturata/non strutturata*, cioè se i dati sono conformi a un certo schema dove i campi hanno un preciso significato (strutturata), e.g., record database, l’informazione è espressa in linguaggio naturale (non strutturata)…

Information Seeking

- *Informazione non strutturata:*

  "**TEHRAN, Iran Sep 1, 2006 (AP)**— Iran underlined its disregard Friday for the U.N. deadline to halt uranium enrichment now expired when its president vowed never to give up its nuclear program and accused the West of misrepresenting Tehran's nuclear activities…"

- *Informazione strutturata:*

Name	Address	Email
Thomas	1989 California St.	thomas@acm.com
Mary	9999 Coyote Hill Rd	mary@mac.com
George	2121 El Camino Real	george@iana.org

Information Seeking

- Esitono anche informazioni parzialmente strutturate, ad esempio…

```
Return-Path: bounce-276592-422906@listserv.nai.com
Received: from mail.dia.uniroma3.it…
From: AVERT_DAT_Release@avertlabs.com
To: gaspare@dia.uniroma3.it
Subject: Avert Labs Dat Release Notification…
Date: Thu, 24 Aug 2006 20:34:08 -0500
MIME-Version: 1.0
Content-Type: text/plain; charset="ISO-8859-1“
Content-Transfer-Encoding: 8bit
Body: The 4837 daily dat files have been released and
are available for download…
```

Information Seeking

Process	Need	Information Source
Information Filtering – IF	Stable & Specific	Dynamic & Unstructured (or Semistructured)
Alerting	Stable & Specific	Dynamic & Structured
Data Mining	Stable & Specific	Stable
Information Retrieval - IR	Dynamic & Specific	Stable & Unstructured
Database Access	Dynamic & Specific	Stable & Structured
Exploration	Broad	Varied

"The State of the Art in Text Filtering" di D.W.Oard. UMUAI 7: 141-178, '97.

Overview

- Motivations & Goals
- Definitions
  - Information
  - Information Overload
  - Information Needs
  - Information Seeking
  - Information Filtering
  - Information Retrieval

Information Filtering

- "In the IF process the user is assumed to be seeking information which addresses a specific long-term interest" (Oard '97)
- IF systems are designed to sort through *large* volumes of *dynamically generated* information and present the user with sources that are likely to satisfy his information needs
- IF systems may either provide these entities directly or it may provide the user with references

Information Filtering

- IF's three subtasks:
  1. collecting the information sources,
  2. detecting useful sources,
  3. displaying the useful sources.

Information Filtering vs Retrieval (Belkin&Croft '92)

- Il processo di IR è importante perchè è alla base di molti Information System, e.g., attuali Web search engine…

  ma molte tecniche IR sono state ri-adattate per il Web perché…

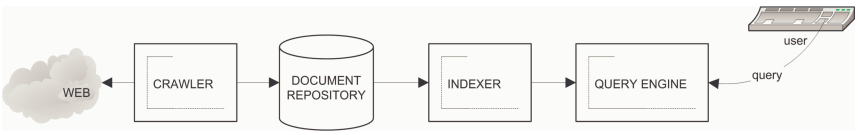
  1. il Web non è una sorgente statica
     - size doubled in 2 years (Lawrence,Giles 1999-'00)
     - 40% pages change weekly,23% `.com` pages daily (Cho,Garcia-Molina'03)
  2. le query utente (derivati dagli information needs) non cambiano sempre nel tempo

Information Filtering vs Retrieval (Belkin&Croft '92)

- Possibile soluzione…
  1. Si aggiornano periodicamente i dati salvati nel motore di ricerca con le modifiche apportate sulle pagine Web (la sorgente non è più considerata statica)
  2. Si suppone che le query di un utente siano scorrelate

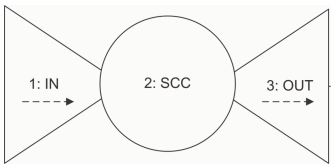
Information Filtering vs Retrieval

- Web Search Engine:
  - Un *crawler* visita il Web scaricando le pagine Web
  - Una copia viene salvata in un *repository* locale
  - Per ogni pagine viene creata una rappresentazione interna (dal *indexer*) usata durante le query



Information Filtering vs Retrieval

- World Wide Web (Broder et al. '00):



  - *SCC* Strongly Connected Component: pages that can reach one another along directed links (27%)
  - *IN & OUT*: pages that can reach the SCC but cannot be reached from it, or that are accessible from the *SCC* but do not link back to it respectively (21% each)
  - the rest cannot be reached and don't reach *SCC* (29%)

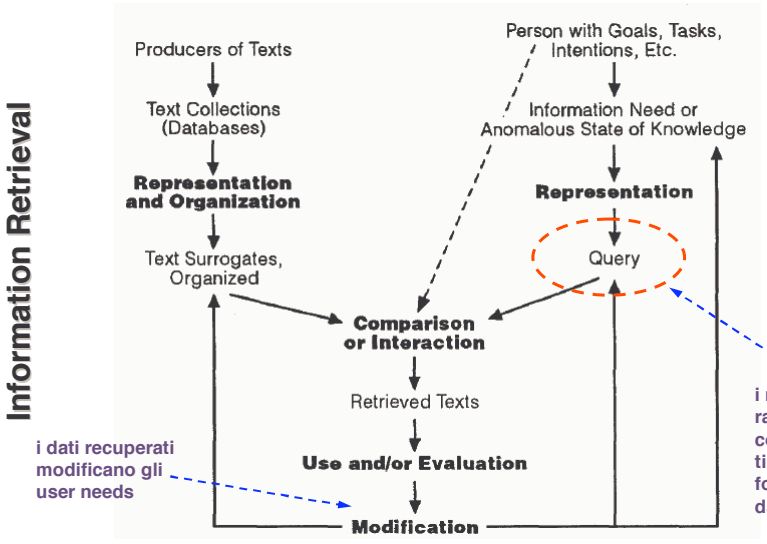
Information Filtering vs Retrieval (Belkin&Croft '92)

- L'IF è basato su *user models* o *profiles* che rappresentano gli interessi a lungo termine; l'utente interagisce più volte col sistema
- Nell'IF le sorgenti sono considerate come *stream* di dati
- L'IF è visto anche come processo di interrogazione di sorgenti esterne (e.g., database) in base agli user needs, vedi *intelligent agents*

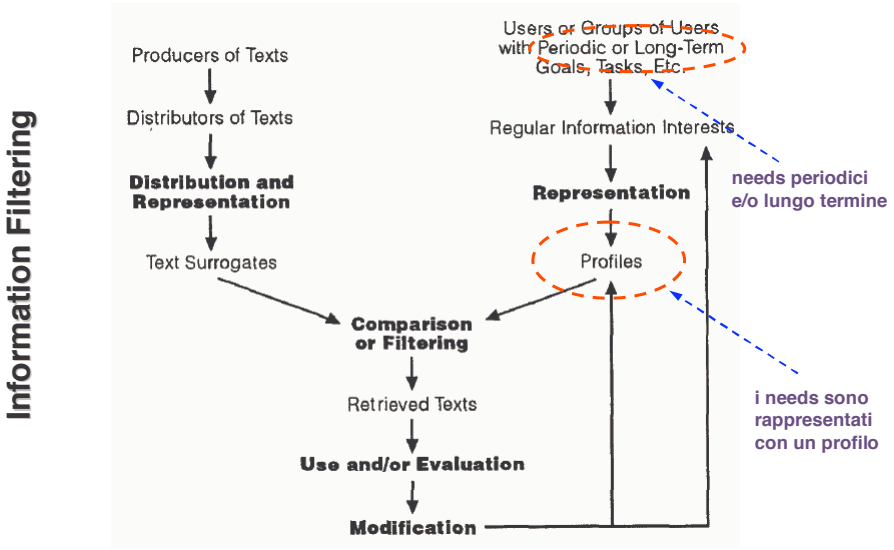
Information Filtering vs Retrieval

- Concettualmente IF rimuove dati da uno stream (e.g., junk emails) mentre IR recupera dati da collezioni…
- Ma IR e IF hanno molto in comune, e.g.:
  - IR's *text routing*: sending relevant incoming data to people
  - IR's *categorization systems:* attach *static* categories to incoming objects

Information Filtering vs Retrieval (Belkin&Croft '92)



Information Filtering vs Retrieval (Belkin&Croft '92)



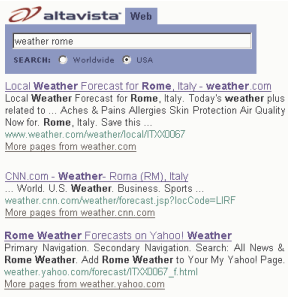
Information Filtering vs Retrieval

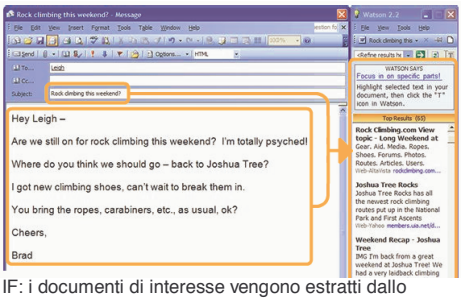
- La presenza di user profiles nel IF implica *rappresentazioni più complesse* dei needs (e documenti) rispetto all'IR
- Nel'IR si studiano tecniche per *ridurre il tempo di recupero* (e.g.,inverted indexes), mentre nel'IF si preferisce aumentare la precisione nei risultati

Information Filtering vs Retrieval

Come vengono generalmente visualizzati i risultati?

IR: l'utente specifica una query e successivamente vengono mostrati i risultati

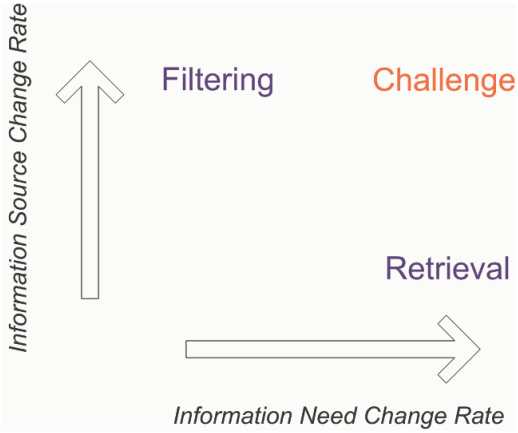




IF: i documenti di interesse vengono estratti dallo stream e proposti all'utente (se ne esistono nuovi oppure sono stati riconosciuti nuovi needs)

Notification popup when new items come in

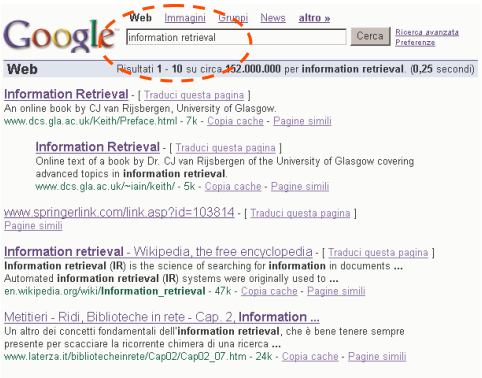
Information Filtering vs Retrieval



Information Retrieval

- Esempio: Motore di Ricerca

data una query (insieme di keywords) si vuole costruire la lista di documenti che la contengono



Overview

- Motivations & Goals
- Definitions
  - Information
  - Information Overload
  - Information Needs
  - Information Seeking
  - Information Filtering
  - Information Retrieval

Information Retrieval

- Che struttura rende possibile tale ricerca molto velocemente?

Docs' content to index

DocId	Text
1	Pease porridge hot, pease porridge cold
2	Pease porridge in the pot
3	Nine days old
4	Some like it hot, some like it cold
5	Some like it in the pot
6	Nine days old

Information Retrieval – Inverted Index

- Quello che ci serve è trovare l'insieme di documenti che contengono una o più parole della query
- Un *indice* (analitico) ci permetterebbe di trovare velocemente tutte le occorrenze all'interno di un documento/collezione

Information Retrieval – Inverted Index

Docs' content to index

DocId	Text
1	Pease porridge hot, pease porridge cold
2	Pease porridge in the pot
3	Nine days old
4	Some like it hot, some like it cold
5	Some like it in the pot
6	Nine days old

Inverted Index

Term	Documents
cold	<2;1,4>
days	<2;3,6>
hot	<2;1,4>
in	<2;2,5>
it	<2;4,5>
…	…

numero occorrenze:2

id doc in cui occorre: 4 e 5

Information Retrieval – Inverted Index

- A cosa serve il numero di occorrenze?
  - Per motivi di efficienza l'indice deve essere molto compatto (37Gbytes per 24M pages 1998 → 1,5Tbytes per 1Billion!)
  - Si tende a memorizzare tutti i record sequenzialmente:

2	1	4	*1*	3	*4*	1	4	2	5	*2*	2	5	…

il numero di occorrenze ci serve per capire quando termina un record

Information Retrieval – Full Inverted Index

Come trattare le *phrase-query ("porridge pease" vs porridge AND pease)?...*

DocId	Text
1	Pease porridge hot, pease porridge cold
2	Pease porridge in the pot
3	Nine days old
4	Some like it hot, some like it cold
5	Some like it in the pot
6	Nine days old

Term	Documents
cold	<2;(1;6),(4;8)>
days	<2;(3;2),(6;2)>
hot	<2;(1;3),(4;4)>
in	<2;(2;3),(5;4)>
it	<2;(4;3,7),(5;3)>
...	...

memorizziamo anche la posizioni delle occorrenze: 3 e 7

Information Retrieval - Matching Booleano

- Per individuare i documenti che soddisfano una query si può utilizzare il *Boolean query processing*:

1. Si prende il 1° termine *t* dalla query
2. Si ricava la lista di doc I_t in cui occorre *t*
3. C ← I_t
4. ∀ restante termine *t'*
   a) si ricava la lista di doc $I_{t'}$ in cui occorre *t'*
   b) ∀doc∈C, se doc∉$I_{t'}$ allora C ← C- {d}
   c) se |C| = 0 allora query non soddisfatta

si può facilmente estendere agli operatori NOT e OR

Information Retrieval – Inverted Index

- Provare a costruire l'inverted index ed "eseguire" la query: *microsoft zune ipod* sui seguenti documenti:

doc #01: Microsoft to Outsource Zune to Toshiba
doc #02: Toshiba shows off Zune
doc #03: Microsoft picks Toshiba to manufacture new music player
doc #04: Toshiba to make Microsoft's Zune media player
doc #05: Microsoft and Toshiba to play a catchy Zune
doc #06: Toshiba to make Zune media player
doc #07: Toshiba to make Microsoft's Zune
doc #08: Microsoft's Zune Picture Clears
doc #09: Microsoft Zune iPod Rival to be Built by Toshiba
doc #10: Toshiba Zune details revealed by the FCC
doc #11: Toshiba to make Zune media player: Microsoft
doc #12: Microsoft: Toshiba to Build IPod Rival
doc #13: Microsoft taps Toshiba to make would-be iPod killer Zune
doc #14: Microsoft says Toshiba to make Zune media player
doc #15: Toshiba to Build Microsoft's Zune

Information Retrieval - Matching Booleano

- Osservazioni:
  1. *Microsoft's ≠ Microsoft?*
  2. si può rendere più efficiente la ricerca notando che *ipod* occorre molto meno di *microsoft…*

     → partire dai termini meno comuni
  3. Se ci sono molti documenti che soddisfano la query è difficile individuare i più attinenti
  4. Non sono possibili matching parziali

Information Retrieval – Ranking

- *Ranking:* assegnazione di un valore di attinenza ad ogni documento proponendo una lista ordinata
- *Probability Ranking Principle*: rank the docs in the collection in the order of their probability of relevance to the user needs, given all the evidence available p(d|E) (Robertson '77)

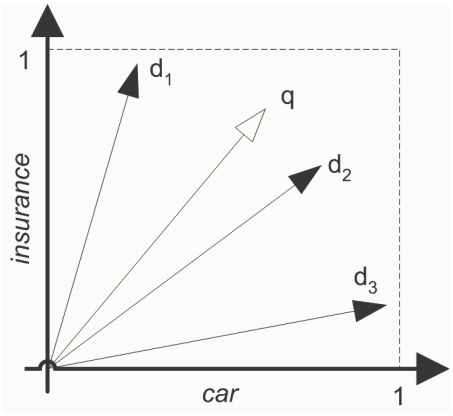
Information Retrieval – VSM-based Matching

- Nel *Vector Space Model* le entità (query e docs) sono rappresentate con vettori in spazi N-dimensionali, dove ogni dimensione è un termine distinto nella collezione
  - Nota: Si assume che i termini siano scorrelati, cioè che che coppie di termini sia ortogonali
- La similarità viene ricavata operando sui vettori, e.g., prodotto scalare

→ Metafora: spatial & semantic proximity

Information Retrieval – VSM-based Matching

- Esempio:
  - spazio 2-dim
  - 3 docs
  - ordine risultati? <d2, d1, d3>

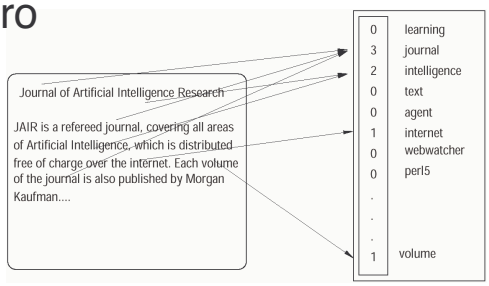


Information Retrieval – VSM-based Matching

- Come determinare i valori degli elementi dei vettori (*weighting*)?
- In base alle occorrenze dei termini nei docs:
  - più un termine occorre in un doc *d*, più è rappresentativo di *d* e perciò riceve peso *w* elevato

Information Retrieval – VSM-based Matching

- Possiamo pensare di rappresentare la query e i documenti come un insieme di parole (rappresentazione *bag of words*) e associare il numero di occorrenze *tf* (*term frequency*) ad ogni elemento del *vettore*

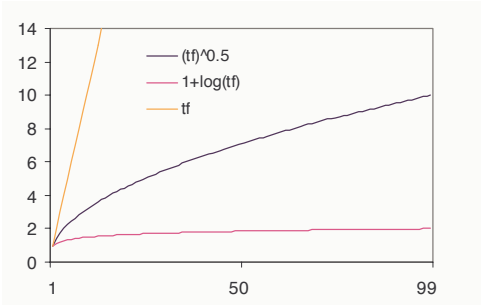


Information Retrieval – VSM-based Matching

- Ma un termine che occorre *3* volte non sempre indica una importanza *x3…*

$f(tf) = (tf)^{0.5}$

$f(tf) = 1+log(tf)$



Information Retrieval – VSM-based Matching

- Ma la rappresentazione basata su tf manifesta problemi…

Esempio:

British Prime Minister Blair held talks on Sunday with President Bush, Russian President Putin and French President Chirac on diplomatic efforts to end the war in Lebanon and Israel… a spokeswoman said the French president had discussed the resolution... a spokesman for Blair said that in the light of Sunday's fighting, the prime minister believed there is an even more urgent need to bring about an end to these hostilities…

Occorrenze:

5 - the
4 - President
3 - to
2 - Sunday
2 - said
2 - Prime
2 - on
2 - minister
2 - in
2 - French
2 - end
2 - Blair
2 - and
2 - an
2 – a
…

Information Retrieval – VSM-based Matching

- Alcuni termini sono termini comuni (e.g., to, the, on, in) mentre altri sono parole “poco importanti” (e.g., end, said, french)
- 2 tecniche risolvono questo problema:
  - si rimuovono i termini comuni con *stop-lists*
  - la frequenza di un termine in una collezione tende ad essere inversamente proporzionale al suo rank (Zipf ’49)

Information Retrieval – VSM-based Matching

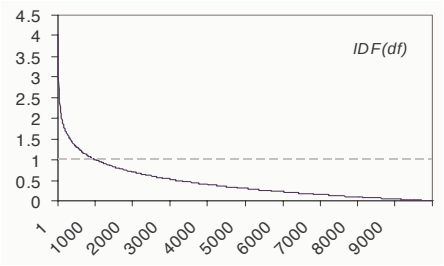
- Una possibile stop-list per l'inglese:

I	in	
a	is	who
about	it	will
an	la	with
are	of	und
as	on	the
at	or	www
be	that	
by	the	
com	this	
de	to	
en	was	
for	what	
from	when	
how	where	

Information Retrieval – VSM-based Matching

- *Inverse Document Frequency* (IDF): misura il grado di *informativeness* di un termine (e.g., *try* vs *insurance*)

$$IDF_j = \log\left(\frac{N}{df_j}\right)$$



N dim collezione
df_j = document frequency di *j*

Information Retrieval – VSM-based Matching

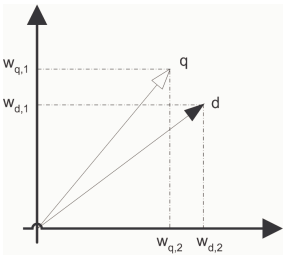
- *Weighting TFxIDF (NTN)*: $d_{x,f} = tf_{x,j} \times \log\left(\frac{N}{df_j}\right)$

Term occurrence	*Document Frequency*	*Normalization (elementi dei vettori)*
N(natural): tf	**N**: df_t	**N**: (no normaliz.)
L(log): 1+log(tf)	**T**: log(N/df)	**C**(cosine): $1/[w_1^2+w_2^2+\ldots+w_n^2]^{\wedge}0.5$
A(augmented): $0.5+[0.5tf / \max_t(tf)]$		

Information Retrieval – VSM-based Matching

- Una volta ottenuti 2 vettori, **d** e **q**, si può ricavarne il *matching*, ad esempio con il prodotto scalare:

$\sum_t w_{q,t} w_{d,t}$



Information Retrieval – VSM-based Matching

- Che risultato si ha con doc1 e doc2?

doc1

A series of bombings on Turkey's Mediterranean coast and in the commercial center of Istanbul has left 27 people wounded, including 10 British tourists, officials report. Meanwhile Prime Minister Blair held talks on Sunday with President Bush, Russian President Putin and French President Chirac on diplomatic efforts to end the war in Lebanon and Israe… Today a six-month amnesty offered by Algeria to Islamic militants on condition of surrender expires on Monday.

doc2

Blair, Bush and Chirac spoke about the war in Lebanon and Israel and a possible resolution

query

Blair Bush Chirac war Lebanon

Information Retrieval – VSM-based Matching

- I documenti più lunghi vengono favoriti perché contengono più termini e il prodotto scalare aumenta
- Solitamente si normalizza il prodotto $Q \cdot D = |Q||D|\cos(QD)$ ottenendo la *cosine rule*:

$$cosine(Q,D) = \frac{\sum_t w_{q,t} w_{d,t}}{\sqrt{\sum_t w_{q,t}^2} \sqrt{\sum_t w_{d,t}^2}}$$

Information Retrieval – VSM-based Matching

- Vantaggi principali VSM
  - basato su spazio vettoriale
  - computazionalmente efficiente
  - molti anni di ricerche disponibili

Information Retrieval – VSM-based Matching

- Problemi del VSM…
  - documenti grandi hanno rappresentazioni che producono basse similarità
  - documenti affini alla query ma con termini diversi non sono recuperati
    (*false negative match*)
  - documenti non affini alla query ma con gli stessi termini vengono recuperati
    (*false positive match*)

Information Filtering vs Alerting

- "Alerting is the database analogue of information filtering, since the only difference between the two is that in an information filtering process it is the information sources (e.g., documents) rather than the information itself which are arriving rapidly" (Oard '97)
- Example: monitoring when mail from a specific user arrives
- The way altering notifies the user can be also used by IF systems

Information Retrieval vs Exploration

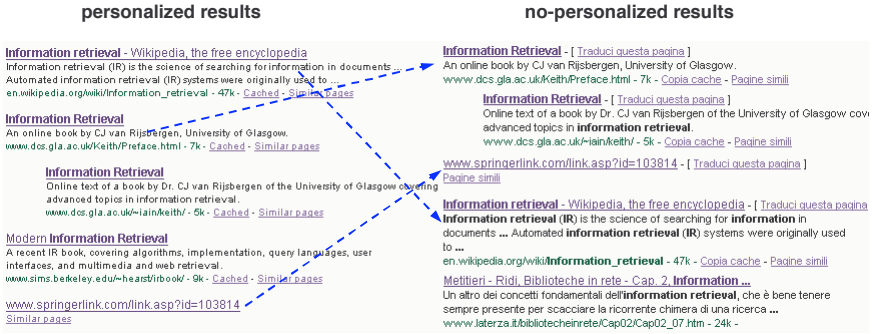
- *searching by query:* directly retrieve documents from huge indexes
  - + quickly retrieve useful information
  - - query formulation
- *searching by surfing* (or *browsing*): in hypertextual environment users analyze Web pages one at a time, surfing through them sequentially, following hyperlinks.
  - + reading and comprehending contents and getting to know the search domain keywords
  - - unable to locate specific pieces of information

Information Filtering vs Retrieval

- Dopo la proposta di molti approcci di IF stand-alone, attualmente si tende ad adattare sistemi di IR incorporando modelli utente; principali vantaggi:
  - a volte si ottengono tempi di recupero simili al caso di IR tradizionale
  - è un approccio adatto anche a sistemi attuali come i search engine
  - si mantengono le stesse interfacce utente

Information Filtering vs Retrieval
Google Personalized

- Esempio con user profile basato su molti papers di IR/IF:



Information Filtering

- Attualmente pochi tool IF disponibili, perché?
  - UM contengono dati sensibili → privacy
  - IF è spesso basato su rappresentazioni complesse dei documenti e user needs → complessità computazionale
  - Nuove interfacce e funzionalità non sempre accettate/sfruttate dagli utenti

Q&A