



Intelligent Web

1 – introduction & definitions

Fabio Gasparetti
gaspare@dia.uniroma3.it

Dipartimento di Informatica e Automazione

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

The Library of Congress >> Switch to Library of Congress Authorities

LIBRARY OF CONGRESS ONLINE CATALOG

New Feature [08/17/06] -- Session Timeout Alert
Please take the new Library of Congress User Survey
Help us improve our services to you!

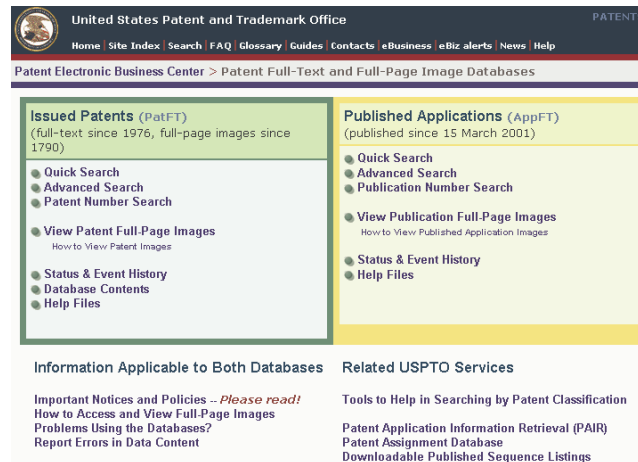
About Displaying and Searching Using Non-Roman Characters
Frequently Asked Questions - Help Table of Contents - What's New

Basic Search	Guided Search	Other Online Catalogs:
<p>Using a fill-in box, search by:</p> <ul style="list-style-type: none"> Title or Author/Creator Subject Call number LCCN, ISSN, or ISBN Keywords <p>Note: Search limits are available only for title and keyword searches.</p>	<p>Using a series of forms and menus:</p> <ul style="list-style-type: none"> Construct keyword searches Restrict all or part of the search to a particular index Combine search words or phrases with Boolean operators <p>Note: Search limits are available for all searches.</p>	<p>Prints and Photographs Online Catalog (PPOC) About - Start Searching</p> <p>Sound Online Inventory & Catalog (SONIC) About - Start Searching</p> <p>Alternative Interface to the LC Online Catalog (Z39.50) About - Start Searching</p> <p>Other Libraries' Catalogs</p>

Information about the images: Two pendentive paintings by Edward J. Holslag are displayed from the Librarian's Room (Librarian's Ceremonial Office) located in the Thomas Jefferson Building of the Library of Congress. On the left, "Efficient clarum studio" (Study, the watchword of fame); on the right, "Dulce ante omnia musae" (The Muses, above all things, delightful).

Online Information Services

www.uspto.gov



United States Patent and Trademark Office

Home Site Index Search FAQ Glossary Guides Contacts eBusiness eBiz alerts News Help

Patent Electronic Business Center > Patent Full-Text and Full-Page Image Databases

Issued Patents (PatFT)
(full-text since 1976, full-page images since 1790)

- Quick Search
- Advanced Search
- Patent Number Search
- View Patent Full-Page Images
How to View Patent Images
- Status & Event History
- Database Contents
- Help Files

Published Applications (AppFT)
(published since 15 March 2001)

- Quick Search
- Advanced Search
- Publication Number Search
- View Publication Full-Page Images
How to View Published Application Images
- Status & Event History
- Help Files

Information Applicable to Both Databases

- Important Notices and Policies -- *Please read!*
- How to Access and View Full-Page Images
- Problems Using the Databases?
- Report Errors in Data Content

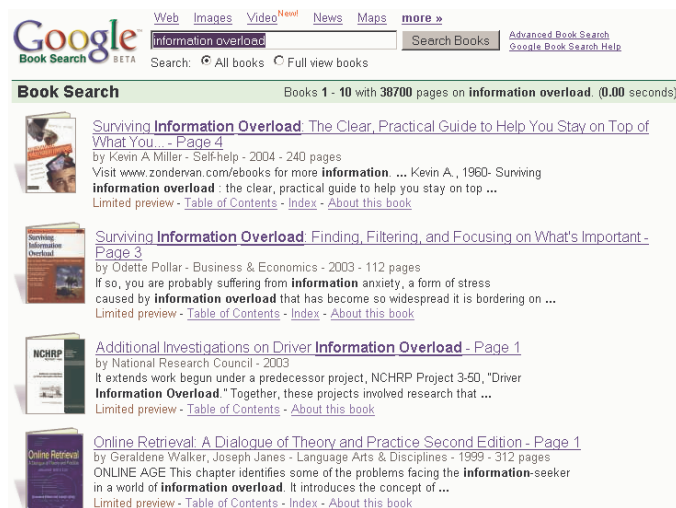
Related USPTO Services

- Tools to Help in Searching by Patent Classification
- Patent Application Information Retrieval (PAIR)
- Patent Assignment Database
- Downloadable Published Sequence Listings

Febbraio 2008

5

Online Information Services



Google Book Search

Web Images Video News Maps more »

Information overload

Search: All books Full view books

Search Books Advanced Book Search Google Book Search Help

Book Search Books 1 - 10 with 38700 pages on information overload. (0.00 seconds)

Surviving Information Overload: The Clear, Practical Guide to Help You Stay on Top of What You... - Page 4
by Kevin A. Miller - Self-help - 2004 - 240 pages
Visit www.zondervan.com/ebooks for more information. ... Kevin A., 1960- Surviving information overload : the clear, practical guide to help you stay on top ...
Limited preview - [Table of Contents](#) - [Index](#) - [About this book](#)

Surviving Information Overload: Finding, Filtering, and Focusing on What's Important - Page 3
by Odette Pollar - Business & Economics - 2003 - 112 pages
If so, you are probably suffering from information anxiety, a form of stress caused by information overload that has become so widespread it is bordering on ...
Limited preview - [Table of Contents](#) - [Index](#) - [About this book](#)

Additional Investigations on Driver Information Overload - Page 1
by National Research Council - 2003
It extends work begun under a predecessor project, NCHRP Project 3-50, "Driver Information Overload." Together, these projects involved research that ...
Limited preview - [Table of Contents](#) - [About this book](#)

Online Retrieval: A Dialogue of Theory and Practice Second Edition - Page 1
by Geraldene Walker, Joseph Janes - Language Arts & Disciplines - 1999 - 312 pages
ONLINE AGE This chapter identifies some of the problems facing the information-seeker in a world of information overload. It introduces the concept of ...
Limited preview - [Table of Contents](#) - [Index](#) - [About this book](#)

Febbraio 2008

6

Online Information Services

Search Engines:

•11.5B indexable pages

The screenshot shows the Altavista search engine interface. The search bar contains 'fettuccine alfredo' and the 'FIND' button is highlighted. Below the search bar, there are filters for 'Worldwide' and 'USA', and 'RESULTS IN: All languages' and 'English, Spanish'. The search results list several links related to 'Fettuccine Alfredo' recipes, including one from e-rcps.com, one from GMA Recipe, one from wchstv.com, one from southernfood.about.com, one from Al Dentel's Fanlisting, and one from Cooks.com.

Febbraio 2008

7

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. $\sim 11 \cdot 10^9$ pagine?
 2. $\sim 50 \cdot 11 \cdot 10^9$ pagine?
 3. $\sim 550 \cdot 11 \cdot 10^9$ pagine?
- *Deep Web* = Web (3) – Indexable Web (a)

Febbraio 2008

8

Online Information Services

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

The screenshot shows the Merriam-Webster Online Dictionary interface. The main content area displays the definition for 'party'. It includes 19 entries found for 'party', with the first 10 listed below. The main entry for 'party' is defined as a noun and a transitive verb. The noun definition includes: 'a person or group taking one side of a question, dispute, or contest', 'a group of persons organized for the purpose of directing the policies of a government', 'a person or group participating in an action or affair', and 'a mountain-climbing party'. The transitive verb definition includes: 'to divide -- more at PART' and 'to divide -- more at PART'. The page also features a sidebar with navigation links and a search bar.

Febbraio 2008

9

Online Information Services

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

The screenshot shows the monster.com job search results page. The page features a search bar at the top with the text 'Refine this search by:'. Below the search bar, there are filters for 'Job Location', 'Job Category', 'Job Title', 'Salary', and 'Status'. The search results are displayed in a table with columns: Date, Job Title, Company, Location, and Job File. The table lists several job openings, including 'DataWarehouse Security Architect' at RS Software, 'Web Developer/Programmer' at Aptimus, Inc., 'Project Manager 5 - Mck-7911009' at Focused HR Solutions, 'Junior/Mid Level Web Developer' at Marcus & Millichap, 'ADMISSIONS DATABASE TECHNICIAN' at UC Hastings College of the Law, 'Unix System Administrator to \$0k+ - Software - Hi-tech' at Accounting Advantage, and 'POMNet Developer/ Engineer II' at Genentech, Inc.

Febbraio 2008

10

Online Information Services

Deep Web: robot excluded pages

The screenshot shows the hotels.com website interface. At the top, there's a navigation bar with links like Home, Flights, Condos, B&B, Vacation Packages, Groups, Deals, Destinations & Interests, and a phone number 800-246-8357. Below this, a search bar displays "Sheraton Los Angeles - Downtown" with a star rating and location details. A "SELECT A ROOM & RATE" button is visible. The main content area includes a "Summary" tab, a "DESCRIPTION" section with a photo of the hotel, and a "Hotel Features" section. The description mentions the hotel's location in the Los Angeles downtown financial district and its amenities like the Grill restaurant and the Enclave bar.

Febbraio 2008

11

Online Information Services

Deep Web: no textual data

The screenshot shows a movie website for "Stranger than Fiction" starring Will Ferrell. The page features a large black exclamation mark graphic, the movie title, and the release date "In Theaters November 10". Below the title, there's a video player showing a scene from the movie with Will Ferrell. A "ENTER THE SITE" button is located at the bottom right of the page.

Febbraio 2008

12

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

INTERNET ARCHIVE
Wayback Machine

Enter Web Address: All

Search Results for Jan 01, 1996 - Sep 01, 2006

1996	1997	1998	1999	2000	2001	2002	2003
0 pages	0 pages	1 pages	12 pages	63 pages	657 pages	159 pages	69 pages
		Dec 02, 1998 *	Jan 17, 1999 *	Feb 29, 2000 *	Jan 19, 2001 *	Jan 23, 2002 *	Feb 02, 2003 *
			Jan 25, 1999 *	Mar 01, 2000 *	Jan 19, 2001 *	Jan 24, 2002 *	Feb 04, 2003 *
			Feb 08, 1999 *	Mar 01, 2000 *	Jan 19, 2001 *	Jan 24, 2002 *	Feb 05, 2003 *
			Feb 09, 1999 *	Mar 02, 2000 *	Jan 19, 2001 *	Feb 06, 2002 *	Feb 08, 2003 *
			Apr 22, 1999 *	Mar 03, 2000 *	Feb 01, 2001 *	Feb 22, 2002 *	Feb 14, 2003 *
			Apr 23, 1999 *	Mar 04, 2000 *	Feb 24, 2001 *	Feb 23, 2002 *	Feb 15, 2003 *
			Apr 27, 1999 *	Apr 08, 2000 *	Feb 26, 2001 *	May 23, 2002 *	Feb 17, 2003 *
			Apr 28, 1999 *	Apr 09, 2000 *	Mar 01, 2001 *	May 25, 2002 *	Mar 19, 2003 *
			May 08, 1999 *	May 10, 2000 *	Mar 01, 2001 *	Jun 02, 2002 *	Mar 24, 2003 *
			Oct 12, 1999 *	May 10, 2000 *	Mar 01, 2001 *	Jun 04, 2002 *	Mar 28, 2003 *
			Nov 06, 1999 *	May 10, 2000 *	Mar 01, 2001 *	Jun 05, 2002 *	Mar 29, 2003 *
			Nov 29, 1999 *	May 10, 2000 *	Mar 02, 2001 *	Jul 02, 2002 *	Apr 02, 2003 *
				May 10, 2000 *	Mar 31, 2001 *	Jul 03, 2002 *	Apr 03, 2003 *
				May 10, 2000 *	Mar 31, 2001 *	Jul 03, 2002 *	Apr 09, 2003 *
				May 10, 2000 *	Apr 01, 2001 *	Jul 04, 2002 *	Apr 21, 2003 *
				May 11, 2000 *	Apr 02, 2001 *	Jul 07, 2002 *	Apr 23, 2003 *
				May 11, 2000 *	Apr 04, 2001 *	Jul 09, 2002 *	Apr 25, 2003 *
				May 11, 2000 *	Apr 04, 2001 *	Jul 20, 2002 *	Apr 25, 2003 *
				May 11, 2000 *	Apr 13, 2001 *	Jul 26, 2002 *	Apr 26, 2003 *
				May 11, 2000 *	Apr 29, 2001 *	Aug 02, 2002 *	May 01, 2003 *
				May 11, 2000 *	Apr 30, 2001 *	Aug 02, 2002 *	May 12, 2003 *
				May 12, 2000 *	May 03, 2001 *	Aug 03, 2002 *	May 27, 2003 *
				May 12, 2000 *	May 04, 2001 *	Aug 04, 2002 *	May 30, 2003 *

Febbraio 2008

13

Online Information Services

Google Web Images Video News Maps more »

Google Search

Top Stories

Iran Defies UN on Enriching Uranium
ABC News - [all 2861 related >](#)

Iranian plane catches fire, killing 80 TV
Reuters uk - [all 115 related >](#)

Donors raise USD 500 million for Palestinians
Ynetnews - [all 63 related >](#)

BBC News | World | UK Edition

Scores killed in Iran plane blaze
Hurricane nears Mexico peninsula
Uganda cash help for war victims

TIME Magazine Online: Top Stories

Making Good on California's Global Warming Gambit
Should Russia Share Blame for the Beslan Massacre?
Why the US Is Holding Its Fire on Iran

Google Videos

Slashdot

Sun Cancels UltraSPARC III+
Internet Not the Social Hinder it Was
Pinto Making a Comeback
Redmond Yawning at Apple-Google Alliance?

Wired News: Top Stories

Netflix Sets Films Free With DVD
Deep Space Wine
Don't Cheat on This Quiz

Techdirt

Canadians Say Please Tax Our Blank CDs Even More
It's Not Whether You Win Or Lose (Your iPod), It's How You Sue Your Friends
Opponents Of Cell Phones On Planes Getting Desperate

CNET News.com

YouTube throws a punch at Facebook
Hackers crack Apple, Microsoft music codes
Web giants lure developers
Photos: Six high-capacity flash players
Realistic face sketching made easy

Tom's Hardware

Reuters: Business News

Alcatel to buy Nortel UMTS unit for \$320 mln
US stock futures edge up, August payrolls watched
Oil steadies above \$70, Iran, Nigeria in focus

CNNMoney.com Latest News

Asian stocks struggle to make gains
Cooking up innovations
Intel mulls cutting 10 percent of workforce

Stock Market

AAPL	67.85	+0.00 (0.00%)
ADBE	32.44	+0.00 (0.00%)
AMD	24.99	+0.00 (0.00%)
AMZN	30.63	+0.00 (0.00%)
CSCO	21.99	+0.00 (0.00%)
EBAY	27.82	+0.00 (0.00%)
GOOG	378.63	+0.00 (0.00%)
IBM	80.97	+0.00 (0.00%)
INTC	19.57	+0.00 (0.00%)
MFE	22.76	+0.00 (0.00%)
MSFT	25.70	+0.00 (0.00%)
NAPS	3.36	+0.00 (0.00%)
ORCL	15.66	+0.00 (0.00%)
RNWK	11.03	+0.00 (0.00%)
SNDK	58.87	+0.00 (0.00%)
SUNW	4.99	+0.00 (0.00%)

Febbraio 2008

14

Online Information Services

The screenshot shows the MY Yahoo! homepage with a navigation bar at the top. Below the navigation bar, there are several sections: 'TV Listings' with a table of programs, 'Weather' for Schenectady, NY, 'Stock Portfolios' with a table of stock prices, 'News' with a list of headlines, and 'Movie Showtimes' for Loew's Rotterdam Square Mall. The page is dated Wednesday, August 4, 10:00 am.

Febbraio 2008

15

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

Febbraio 2008

16

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 factual¹ 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. | <i>Omission</i> : failing to process some of the inputs |
| | 2. | <i>Error</i> : processing the information incorrectly in some way |
| | 3. | <i>Escaping</i> : giving up the burden of attending the inputs entirely |
| can be dysfunctional if not based on well-considered priorities | 4. | <i>Queuing</i> : delaying the processing of some information with the intention of catching up later |
| | 5. | <i>Approximation</i> : lowering standards of discrimination by being less precise in categorizing inputs and responses |
| | 6. | <i>Multiple Channels</i> : 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)



Information Needs

1. **visceral**: "conscious or even unconscious need for information... a vague sort of dissatisfaction...probably inexpressible 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

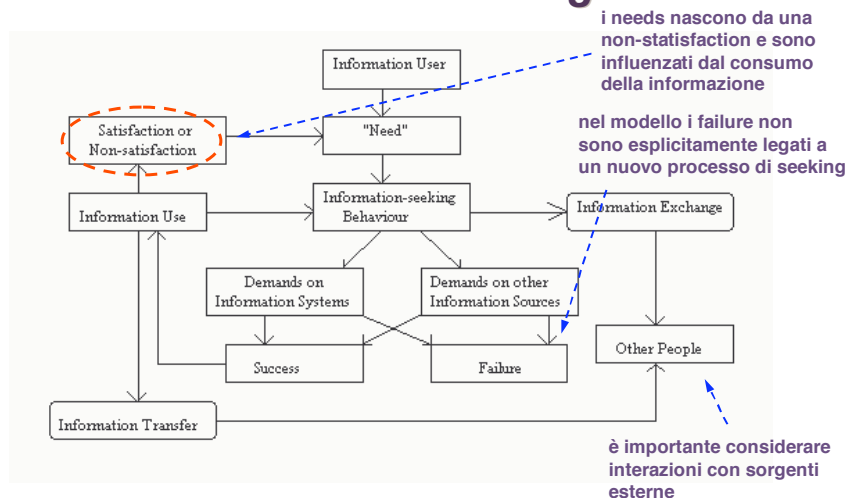
- “It takes place when a person recognize a *gap in their knowledge* that may motivate that person to acquire new information”
(Zerbinos '90)
- “a process in which humans purposefully engage in order *to change their state of knowledge* closely related to learning and *problem solving*” (Marchionini '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

□ Esistono 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
<i>Information Filtering – IF</i>	Stable & Specific	Dynamic & Unstructured (or Semistructured)
<i>Alerting</i>	Stable & Specific	Dynamic & Structured
<i>Data Mining</i>	Stable & Specific	Stable
<i>Information Retrieval - IR</i>	Dynamic & Specific	Stable & Unstructured
<i>Database Access</i>	Dynamic & Specific	Stable & Structured
<i>Exploration</i>	Broad	Varied

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

Overview

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

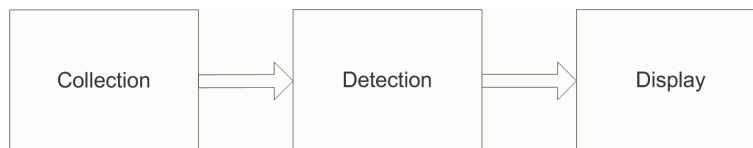
Information Filtering

- “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

The screenshot shows the Blinkx TV website interface. At the top, there's a navigation bar with links like 'personal TV channels', 'blinkx.com', 'upload video', and 'help'. Below this is a search bar with the query 'Pluto loses planet status'. The search results are displayed in a list format. Each result includes a thumbnail, a title, a brief description, and a date. A red dashed circle highlights one of the search results, which is titled 'Cougars of Big Cat Rescue'. Below this, there's a video player showing a clip of a man speaking at a podium. Annotations with arrows point to the search results, labeling them as 'reference (url)' and 'preview'. The video player has a 'Switch to Movie Mode' button. The sidebar on the left lists various news sources like CNN, Reuters, BBC News, etc.

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

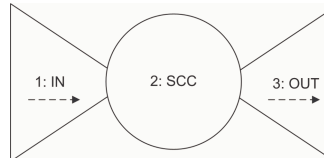
Information Filtering vs Retrieval

- 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%)

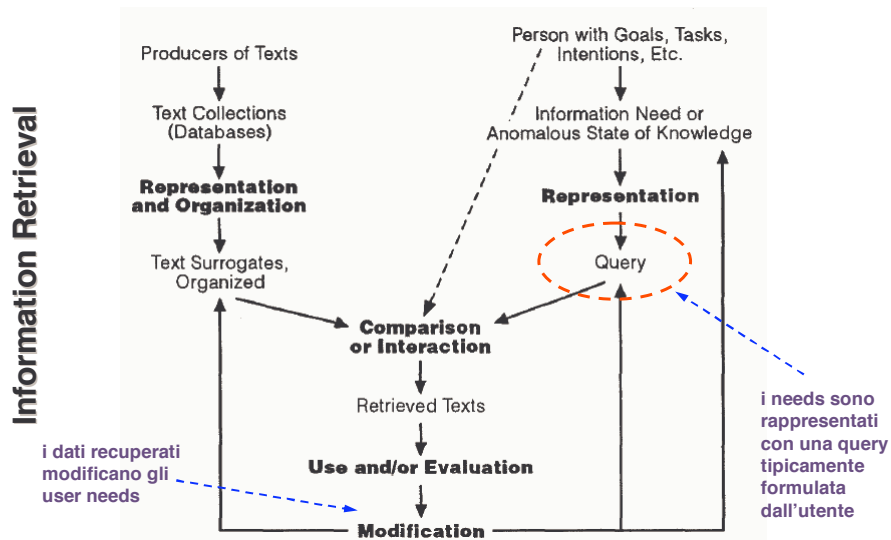
Information Filtering vs Retrieval (Belkin&Croft '92)

- 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*

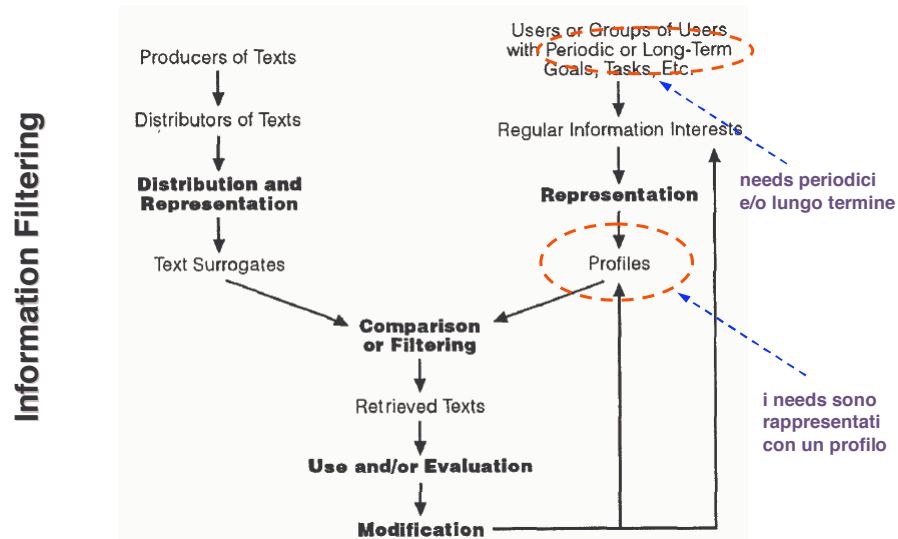
Information Filtering vs Retrieval

- 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)



Febbraio 2008

55

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

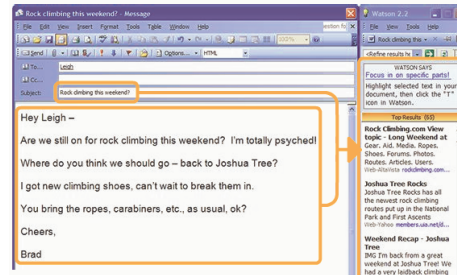
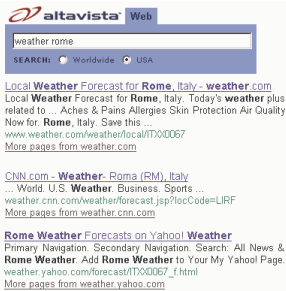
Febbraio 2008

56

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)

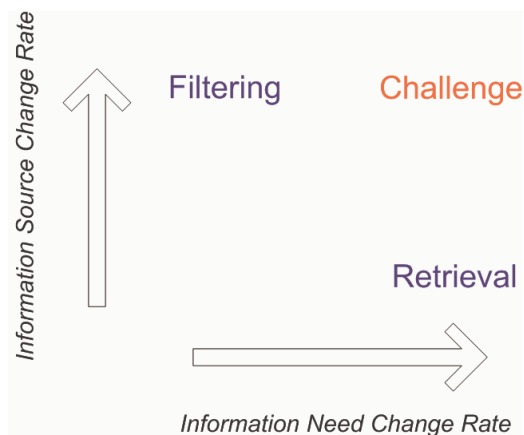


Notification popup when new items come in

Febbraio 2008

57

Information Filtering vs Retrieval



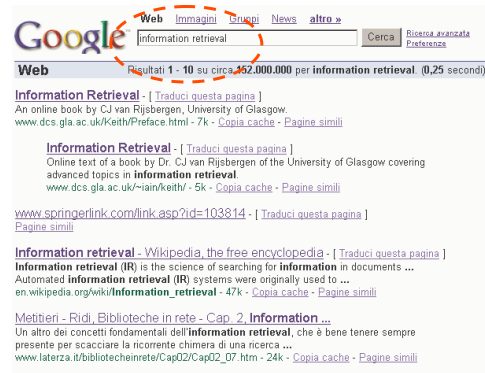
Febbraio 2008

58

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

HyperClass algorithm, 192
 hyperlink induced topic search. See HITS
 hyperlinks, 17
 analysis, 9–10
 communities, 73
 extraction, 25–26
 generated from templates in navigation bars, 221
 nepotistic, 220
 repeated expansion of, avoiding, 29
 scanning for, 19
 as similarity indicators, 118
 hyperplane, 165–166
 defined, 164
 optimized, 166
 hypertext, 1, 17
 data structure, 5
 defined, 1
 features, 9
 supervised learning for, 169–173
 Xanadu system, 1

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. $C \leftarrow I_t$
 4. \forall restante termine t'
 - a) si ricava la lista di doc $I_{t'}$ in cui occorre t'
 - b) $\forall \text{doc} \in C$, se $\text{doc} \notin I_{t'}$ allora $C \leftarrow C - \{\text{doc}\}$
 - 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* \neq *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

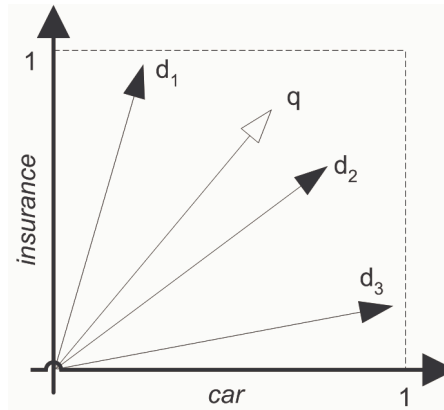
- *Ranking*: assegnazione di un valore di attinenza ad ogni documento proponendo una lista ordinata
- *Probability Ranking Principle*: rank the docs in the collection in the order of their probability of relevance to the user needs, given all the evidence available $p(d|E)$ (Robertson '77)

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 le 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>

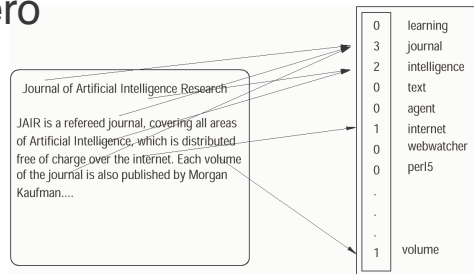


Information Retrieval – VSM-based Matching

- 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



Febbraio 2008

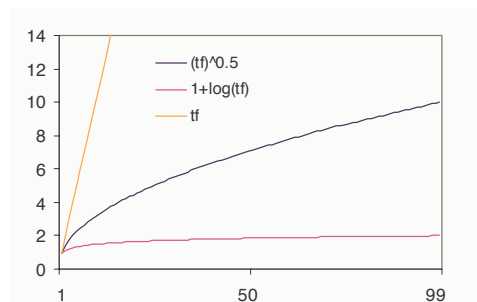
73

Information Retrieval – VSM-based Matching

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

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

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



Febbraio 2008

74

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

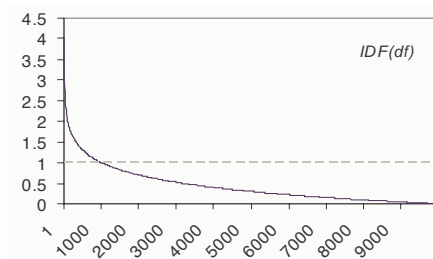
Information Retrieval – VSM-based Matching

- *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

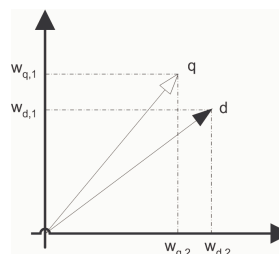
- **Weighting TFxIDF (NTN):** $d_{x,f} = tf_{x,j} \times \log \left(\frac{N}{df_j} \right)$

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): $1/[w_1^2 + w_2^2 + \dots + w_n^2]^{0.5}$
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 Israel... 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*:

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

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 alerting 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:

personalized results

Information retrieval - Wikipedia, the free encyclopedia
Information retrieval (IR) is the science of searching for information in documents ... Automated information retrieval (IR) systems were originally used to ...
en.wikipedia.org/wiki/Information_retrieval - 47k - [Cached](#) - [Similar pages](#)

Information Retrieval
An online book by CJ van Rijsbergen, University of Glasgow.
www.dcs.gla.ac.uk/Keith/Preface.html - 7k - [Cached](#) - [Similar pages](#)

Information Retrieval
Online text of a book by Dr. CJ van Rijsbergen of the University of Glasgow covering advanced topics in information retrieval.
www.dcs.gla.ac.uk/~iain/keith/ - 5k - [Cached](#) - [Similar pages](#)

Modern Information Retrieval
A recent IR book, covering algorithms, implementation, query languages, user interfaces, and multimedia and web retrieval.
www.sims.berkeley.edu/~heast/irbook/ - 9k - [Cached](#) - [Similar pages](#)

www.springerlink.com/link.asp?id=103814
[Similar pages](#)

no-personalized results

Information Retrieval - [[Traduci questa pagina](#)]
An online book by CJ van Rijsbergen, University of Glasgow.
www.dcs.gla.ac.uk/Keith/Preface.html - 7k - [Cached](#) - [Pagina simili](#)

Information Retrieval - [[Traduci questa pagina](#)]
Online text of a book by Dr. CJ van Rijsbergen of the University of Glasgow covering advanced topics in information retrieval.
www.dcs.gla.ac.uk/~iain/keith/ - 5k - [Cached](#) - [Pagina simili](#)

www.springerlink.com/link.asp?id=103814 - [[Traduci questa pagina](#)]
[Pagina simili](#)

Information retrieval - Wikipedia, the free encyclopedia - [[Traduci questa pagina](#)]
Information retrieval (IR) is the science of searching for information in documents ... Automated information retrieval (IR) systems were originally used to ...
en.wikipedia.org/wiki/Information_retrieval - 47k - [Cached](#) - [Pagina simili](#)

Mettitieri - Ridi. Biblioteche in rete - Cap. 2. **Information ...**
Un altro dei concetti fondamentali dell'information retrieval, che è bene tenere sempre presente per scacciare la ricorrente chimera di una ricerca ...
www.laterza.it/bibliotecheinrete/Cap02/Cap02_07.htm - 24k -

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