

# *INTRODUZIONE ALL'INFORMATION RETRIEVAL*

---

CORSO DI  
*WEB MINING & RETRIEVAL, A.A. 2013-14*

Roberto Basili

# Outline

- *Accesso e Ricerca delle informazioni distribuite*
- Il processo di base dell'IR
  - Rilevanza
- Applicazioni dell'IR:
  - Classification
  - Inf. Filtering & Routing
  - Text Clustering
  - Inf. Extraction, Question Answering
- Web search

# PARTE II: INFORMATION RETRIEVAL

---

Introduzione

# Reperimento della Informazione

- Se la memorizzazione (mediante dispositivi di memoria di massa) e' massiva (testi, immagini, suoni, ...) si pone il problema di **ricercare** tale informazione
- Il livello di astrazione consentito dai Sistemi Operativi (File System) e' solo un primo livello:
  - e' insufficiente in molti casi (ad es. anagrafica)
  - non e' ottimale (riguardo alla velocità della ricerca)

# Reperimento della Informazione

- *Ricerca* in generale significa
  - definire i propri bisogni informativi
  - memorizzare i risultati
  - raffinare la propria selezione
    - ridefinire i requisiti informativi
    - “navigare” attraverso i dati trovati
    - elaborare, cioè combinare i dati di diverse ricerche

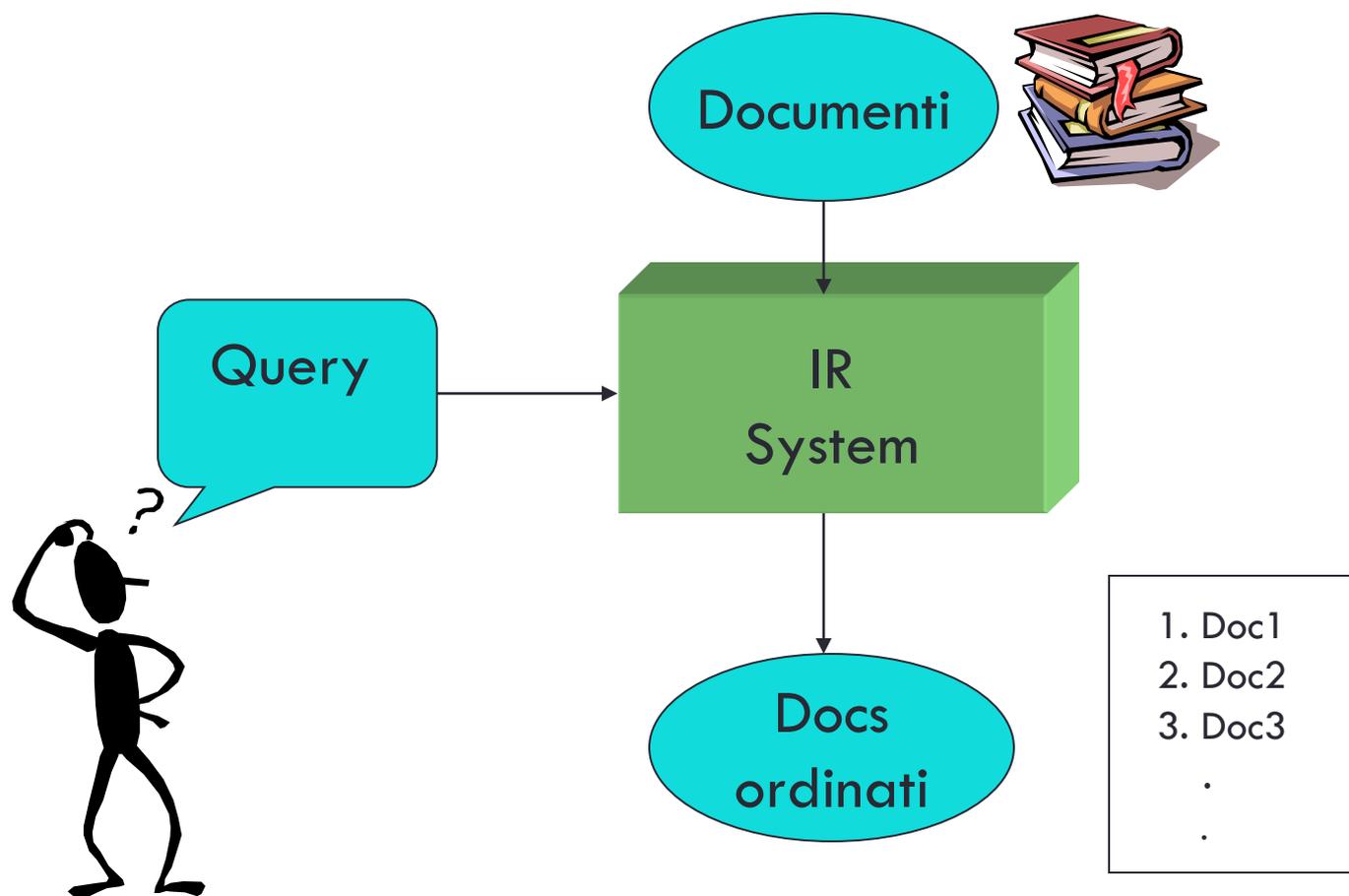
# Reperimento della Informazione (IR): Strumenti

- determinano (a priori) una strutturazione dell'informazione da ricercare che
  - rifletta il linguaggio di interrogazione
  - rifletta la natura (cioè il contenuto) dei dati da ricercare (vd. JPEG, BMP, WAV!!)
  - faciliti le operazioni interne di ricerca
- forniscono un linguaggio per la definizione dei bisogni informativi, detto linguaggio di interrogazione (*Query language*)

# Tipico processo di IR

- Dati:
  - Una **collezione di documenti** in linguaggio naturale.
  - Una **interrogazione** utente (in genere una stringa di testo)
- Trovare:
  - Un **elenco ordinato di documenti rilevanti** per la interrogazione (l'ordinamento e' decrescente)

# Il processo di IR



# Rilevanza (Attinenza)

- La **attinenza** di un documento ad una interrogazione (query) e' **soggettiva** e dipende da:
  - appartenenza ad un **campo semantico** (soggetto)
  - **puntualità** (essere recente ed al momento giusto)
  - **autorevolezza** (provenienza sicura)
  - **Pertinenza agli obiettivi dell'utente** ed al suo utilizzo dell'informazione

# Relevance

Why has information science emerged on its own and not as a part of librarianship or documentation, which would be most logical? It has to do with relevance. ...to be effective, scientific communication ...has to deal not with any old kind of information but with relevant information.

[Sar75, pages 323-324]

... the topic of relevance, acknowledged as the most fundamental and much debated concern for information science... Early on, information scientists recognized that the concept of relevance was integral to information system design, development and evaluation. However, there was little agreement as to the exact nature of relevance and even less that it could be operationalized in systems or for the evaluation of systems. ...this lack of agreement continues to an extent at the present.

[Fro94, page 124]

da Stefano Mizzaro, "*Relevance: The Whole History*" in Journal of the American Society of Information Science, volume 48, (9), 810-832, 1997, URL = "[citeseer.ist.psu.edu/mizzaro96relevance.html](http://citeseer.ist.psu.edu/mizzaro96relevance.html)"

# Keyword (Parole chiave)

- Una *keyword* e' costituita di una o piu' parole
  - *rugby, Scozia, Italia*
  - *6 Nazioni, Istituto di Fisica Matematica*
- Costituiscono la nozione piu' semplice di *attinenza*, i.e.
  - *Occorrenza letterale nel testo*
- Unico compromesso:
  - Le parole definite come keyword debbono apparire frequentemente nel documento, *indipendentemente dal loro ordine* (bag of words).

# Limitazioni delle *keywords*

## Variabilità

- (*Silenzio*) non vengono trovati documenti che includano (solo) termini sinonimi
  - “*imposta*” vs. “*tassa*”, “*basket*” vs. “*pallacanestro*”
  - “*Stati Uniti*” vs. “*USA*”

## Ambiguità

- (*Rumore*) vengono ritrovati documenti che includono anche termini ambigui
  - “*imposta*” (finestra vs. tassa)
  - “*Apple*” (company vs. frutta)
  - “*operare*” (in mercato vs. chirurgia)
  - “*Jaguar*” (macchina vs. software)

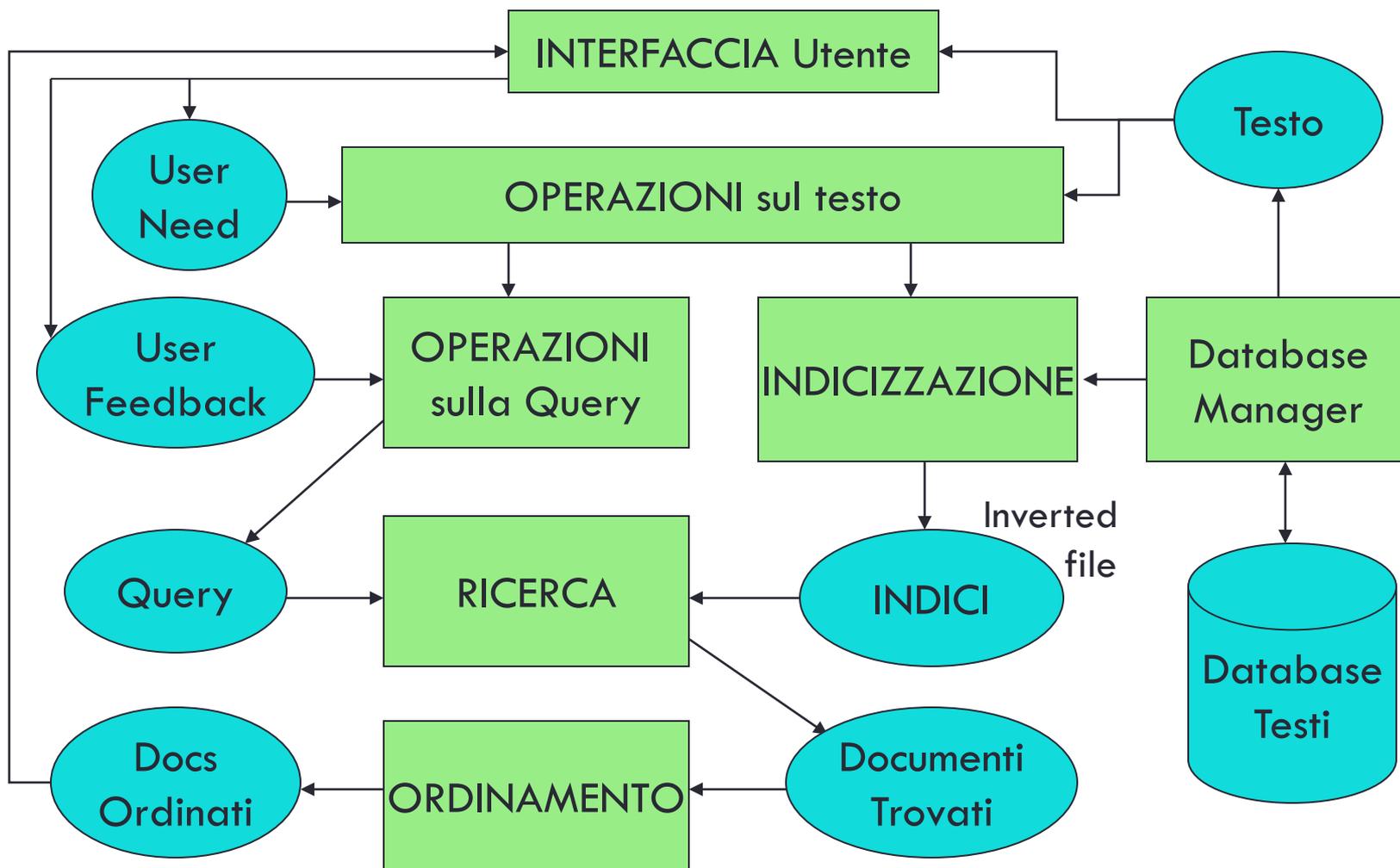
## ... oltre le *keywords*

- Le tecniche keyword-based e soprattutto ...
- le estensioni e gli sviluppi recenti più espressivi (sensi, nomi propri, n-grams, ...)
- Modelli collaborativi
  - Social media
  - Collaborative filtering
- Apprendimento automatico per il sostegno a
  - Personalizzazione
  - Sviluppo su larga scala (Cloud Computing)
- Alla base di tali estensioni ci sono sempre legami con altri paradigmi: Intelligenza Artificiale, Web Semantico e Ingegneria del software

# IR intelligente

- Rendere sensibile il sistema al *significato* delle parole
  - Es. *imposta*<sub>tassa</sub>, *imposta*<sub>finestra</sub>
- Considerare l'*ordinamento* delle parole della interrogazione
  - Es. *computer science* vs. *science and computers*
- Considerare l'utente sulla base di un "*feedback*" esplicito o implicito
- Considerare informazioni sulla sorgente (ad es. *autorità/affidabilità delle fonti*)

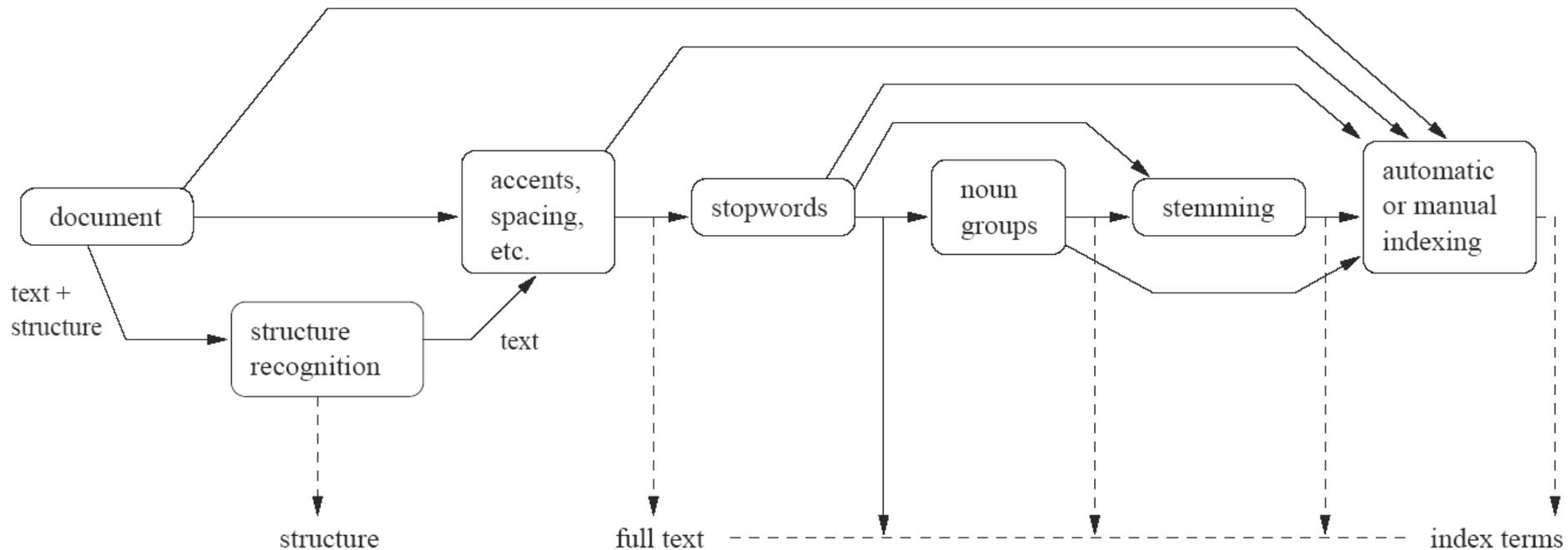
# Architettura di un sistema di IR



# Sistemi di IR: Componenti

- Operazioni sui Testi
  - Selezione degli indici.
  - Rimozione delle Stopword
  - Stemming/Lemmatizzazione

# Operazioni sui Testi.



# Sistemi di IR: Componenti (2)

- **INDICIZZAZIONE**

- Costruisce l'indice inverso:  
→ riferimenti ai documenti

parole

- **RICERCA**: trova i documenti che includono un elemento della interrogazione (usando l'indice inverso)
- **ORDINAMENTO** dei documenti trovati secondo i *valori di attinenza*.

# Sistemi di IR: Componenti (3)

- **Interfacce utente: gestiscono le interazioni**
  - Inserimento interrogazione e visualizzazione dei documenti.
  - Relevance feedback.
  - Visualizzazione dei risultati.
- **Operazioni sulla Query: trasformano la query per migliorare le prestazioni:**
  - Espansione (*Query expansion*), per es. mediante un *thesaurus*.
  - Trasformazione (pesatura) mediante relevance feedback.

## IR: Ulteriori *task*

- Categorizzazione Automatica di Documenti
- *Information filtering (spam filtering)*
- *Information routing*
- *Document clustering*
- *Recommending information or products*
- *Information extraction and Summarisation*
- *Question answering*
- *Opinion Mining*

# IR: Contigua' Disciplinare

- *Database Management*
- *Library and Information Science*
- *Artificial Intelligence*
- *Natural Language Processing*
- *Machine Learning*

# Machine Learning

- Focus sullo sviluppo di sistemi software che migliorano le proprie prestazioni tramite l'esperienza.
- Classificazione Automatica mediante apprendimento supervisionato da esempi (*supervised learning*).
- Metodi automatici di *clustering* di documenti in classi significative (*unsupervised learning for KM*).

# Machine Learning: direzioni verso l'IR

- **Categorizzazione** dei Testi
  - Classificazione Automatica Gerarchica (es. Yahoo).
  - Filtering/Routing/Reccomendation Adattivi
  - Automated *spam filtering*.
- **Clustering** dei Testi
  - Clustering dei risultato di *IR queries*.
  - Sviluppo automatico di gerarchie di classi (Yahoo).
- **Learning to Rank**
- Apprendimento Automatico per l'**Information Extraction**
- **Text Mining**
- Analisi dei dati del Web 2.0 (**Social Web Mining**)

# Text Clustering: Vivisimo

Clusty Search » Roberto Basili - Mozilla Firefox

File Modifica Visualizza Cronologia Segnalibri Strumenti ?

http://clusty.com/search?v%3afile=viv\_1101%4019%3aTEffVk&v%3aframe=list&v%3astate=(root(N840))|root&id=N387&action=list&sw=

Più visitati Corso: Basi di dati FOLIE Home Tree Kernels in SVM-li... Calls EMEROTECA GEMS09 SRL Demo FOLIE REQUIRELogin Review

Today's Paper - ... Rada Mihalcea: ... SSL Tutorial: AC... Facoltà di Ingeg... http://a...ad.html Processing Robu... START RANL

web news images wikipedia blogs jobs more »

**Clusty** Roberto Basili Search advanced preferences

clusters sources sites remix

All Results (195)

- ➔ Alessandro Moschitti (36)
- ➔ Daniele Pighin (9)
- ➔ Text Classification (8)
- ➔ Roberto Basili, Marco Cammisa (6)
- Italy (4)
- Roberto Basili, Diego De Cao, Danilo Croce, Bonaventura (3)
- W05-0601 (2)
- Structural Representations (2)
- Alessandro Moschitti, Silvia Quarteroni, Roberto Basili And Suresh (2)
- Other Topics (4)

Top 195 results of at least 1,690 retrieved for the query **Roberto Basili** (details)

Did you mean: [robert basili](#)

- [Roberto Basili's Home Page](#)

Thanks Department to Round the Bend Wizards: **Roberto Basili** Home. My life in the bush of ghosts ... Since May Science of ...  
ai-nlp.info.uniroma2.it/basili - [cache] - Live, Ask
- [Roberto Basili](#)

advertisement. Overview. STARMeter: Down 12% in popularity this week. See rank & trends on IMDbPro. IMDb Res photos to this ...  
www.imdb.com/name/nm0059890 - [cache] - Live
- [BibSonomy :: user :: jamesh](#)

**Roberto Basili** and Marco Cammisa and Alessandro Moschitti AI\*IA, page290-302 ...  
www.bibsonomy.org/user/jamesh - [cache] - Gigablast
- [hpsg-l mailing list: ECML'98 TANLPS Workshop: First Call for Pa](#)

**Roberto Basili** (basili@info.utovrm.it) Mon, 05 Jan 1998 13:51:29 +0100. Next message: Paul Buitelaar ... Apologi  
hpsg.stanford.edu/hpsg-l/1998/0000.html - [cache] - Live, Ask

# WSD

Clusty Search » Word Sense Disambiguation - Mozilla Firefox

File Modifica Visualizza Cronologia Segnalibri Strumenti ?

http://clusty.com/search?input-form=clusty-simple&v%3Asources=webplus&query=Word+Sense+Disambiguation

Più visitati Corso: Basi di dati FOLIE Home Tree Kernels in SVM-li... Calls EMEROTECA GEMS09 SRL Demo FOLIE REQUIRELogin Review

Today's Paper - ... Rada Mihalcea: ... SSL Tutorial: AC... Facoltà di Ingeg... http://a...ad.html Processing Robu... START RANL

web news images wikipedia blogs jobs more »

**Clusty** Word Sense Disambiguation Search [advanced preferences](#)

clusters sources sites

All Results (182) remix

- + Language, Natural (30)
- + Semantic (22)
- + Computational, Linguistics (19)
- + Supervised (16)
- + Biomedical (10)
- + WordNet (12)
- + Information Retrieval (9)
- + Translation, Machine (10)
- + Methods For Word Sense Disambiguation (6)
- + Book (5)

[more](#) | [all clusters](#)

find in clusters:

Top 182 results of at least 46,500 retrieved for the query **Word Sense Disambiguation** ([details](#))

- Word Sense Disambiguation - Book Website**    **TOPICS**  
Companion web site for the **WSD** book, edited by Eneko Agirre and Phil Elvén, June 2006  
[www.wsdbook.org](http://www.wsdbook.org) - [cache] - Live, Ask, Gigablast
- Word Sense Disambiguation (WSD) Test Collection**     
**Word sense ambiguity** is a pervasive characteristic of natural language. For example, the **word** "cold" has several senses...  
[wsd.nlm.nih.gov](http://wsd.nlm.nih.gov) - [cache] - Live, Ask, Gigablast
- Word Sense Disambiguation**     
Evaluating **Word Sense Disambiguation** Systems There are now many computer programs for automatically determining the sense of a word in context (Word Sense Disambiguation) and we have organized a workshop to evaluate the strengths...  
[www.itri.brighton.ac.uk/events/senseval](http://www.itri.brighton.ac.uk/events/senseval) - [cache] - Live, Ask, Gigablast
- Senseval web page**     
There are now many computer programs for automatically determining the **sense of a word** in context (**Word Sense Disambiguation**) and we have organized a workshop to evaluate the strengths...  
[www.senseval.org](http://www.senseval.org) - [cache] - Ask, Gigablast
- Word sense disambiguation - ACLWiki**     
**Word Sense Disambiguation (WSD)** is the ability of software to distinguish what sense of a word is being used in a given context.

# Sommario

- Perché l'IR è importante
- Cos'è l'IR
- Come funziona un sistema generico di IR
- Cosa significa IR "intelligente"
- Quali sono le relazioni di questa tecnologia con altre aree della CS