INTRODUZIONE ALL'INFORMATION RETRIEVAL

CORSO DI
WEB MINING & RETRIEVAL, A.A. 2015-16

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

Dati, informazioni, evidenze e conoscenza

Reperimento della Informazione

- Se la memorizzazione (mediante dispositivi di memoria di massa) e' massiva (testi, immagini, suoni, …) si pone il problema di localizzare, quindi **ricercare, selezionare e recuperare,** 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

- *Ricercare* 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

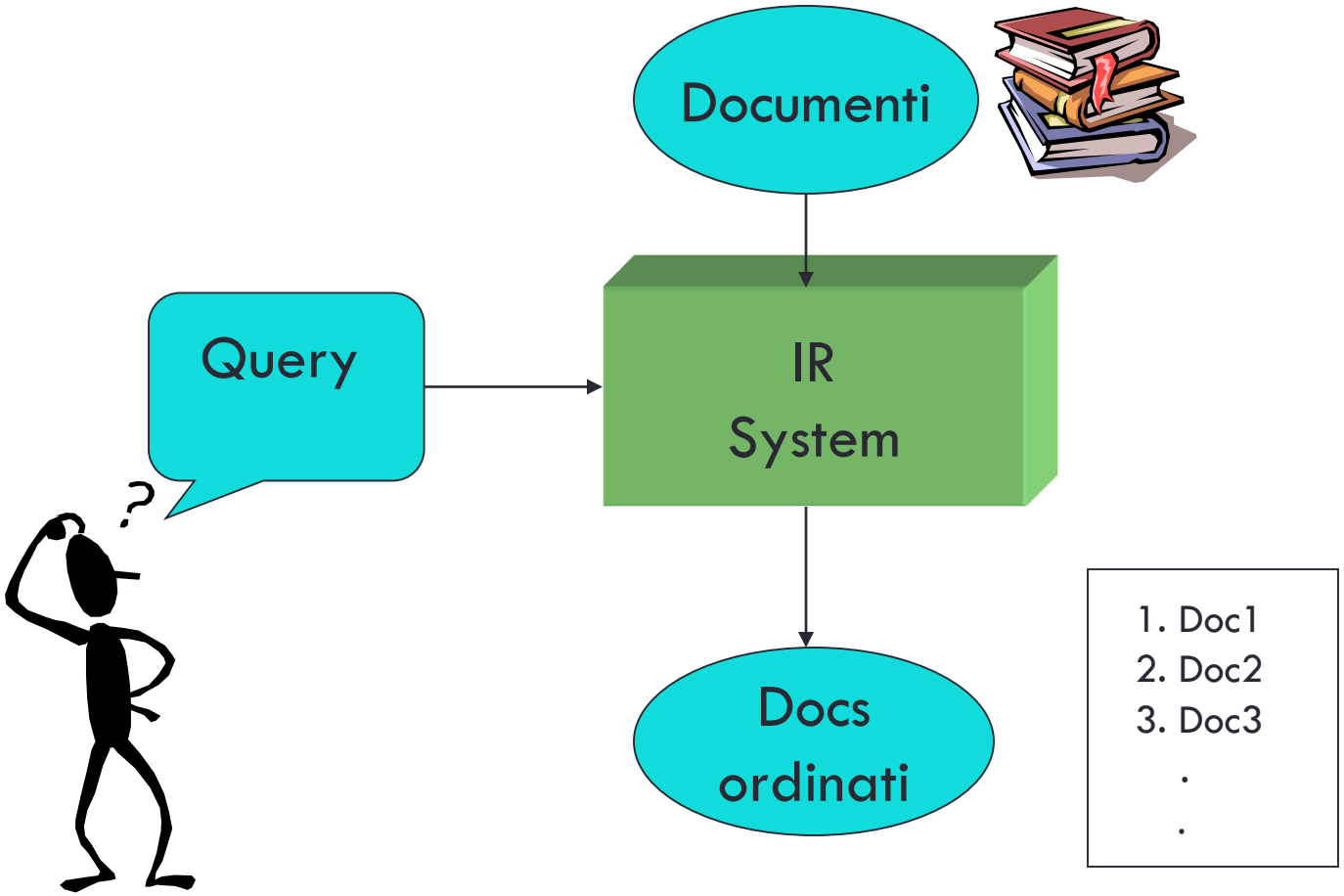
Sistemi per il Reperimento della Informazione (IR Systems)

- 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 obbiettivi 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`"

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*

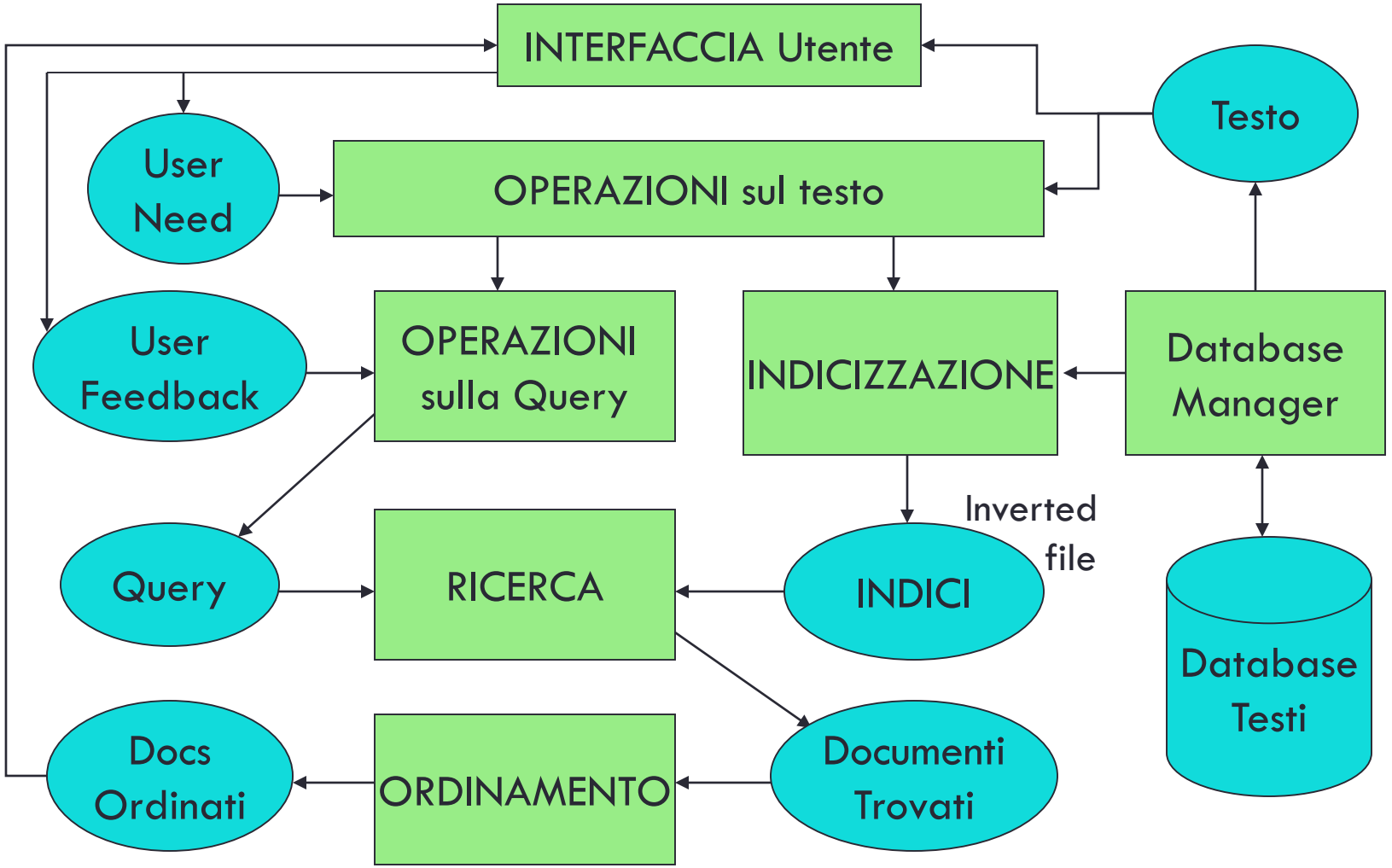
- Le tecniche keyword-based e soprattutto …
- le estensioni e gli sviluppi recenti più espressivi (sensi, nomi propri, n-grams, …)
- Modelli collaborativi
  - Social media
  - Collaborative filtering
- Apprendimento automatico per il sostegno a
  - Personalizzazione
  - Sviluppo su larga scala (Cloud 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
  - alla **sintassi** delle interrogazioni
    - Es. *computer science* vs. *science* and *computers*
  - al **significato** delle parole
    - Es. *impostal*$_{\text{tassa}}$ vs. *impostal*$_{\text{finestra}}$
- Considerare il “feedback” esplicito o implicito ricevuto dall’utente

- Considerare informazioni sulla sorgente (ad es. autorità/affidabilità delle fonti)
- Considerare la comunità dei documenti e della popolazione utente per ottimizzare il processo di IR

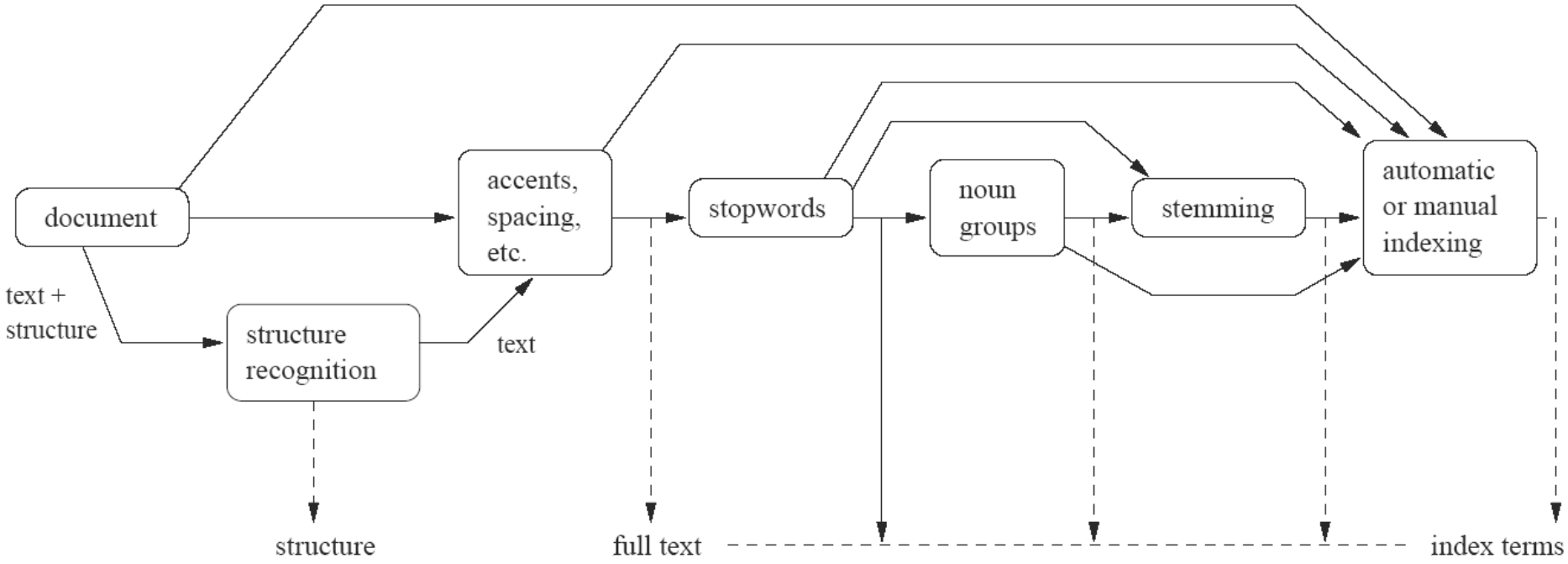
Architettura di un sistema di IR



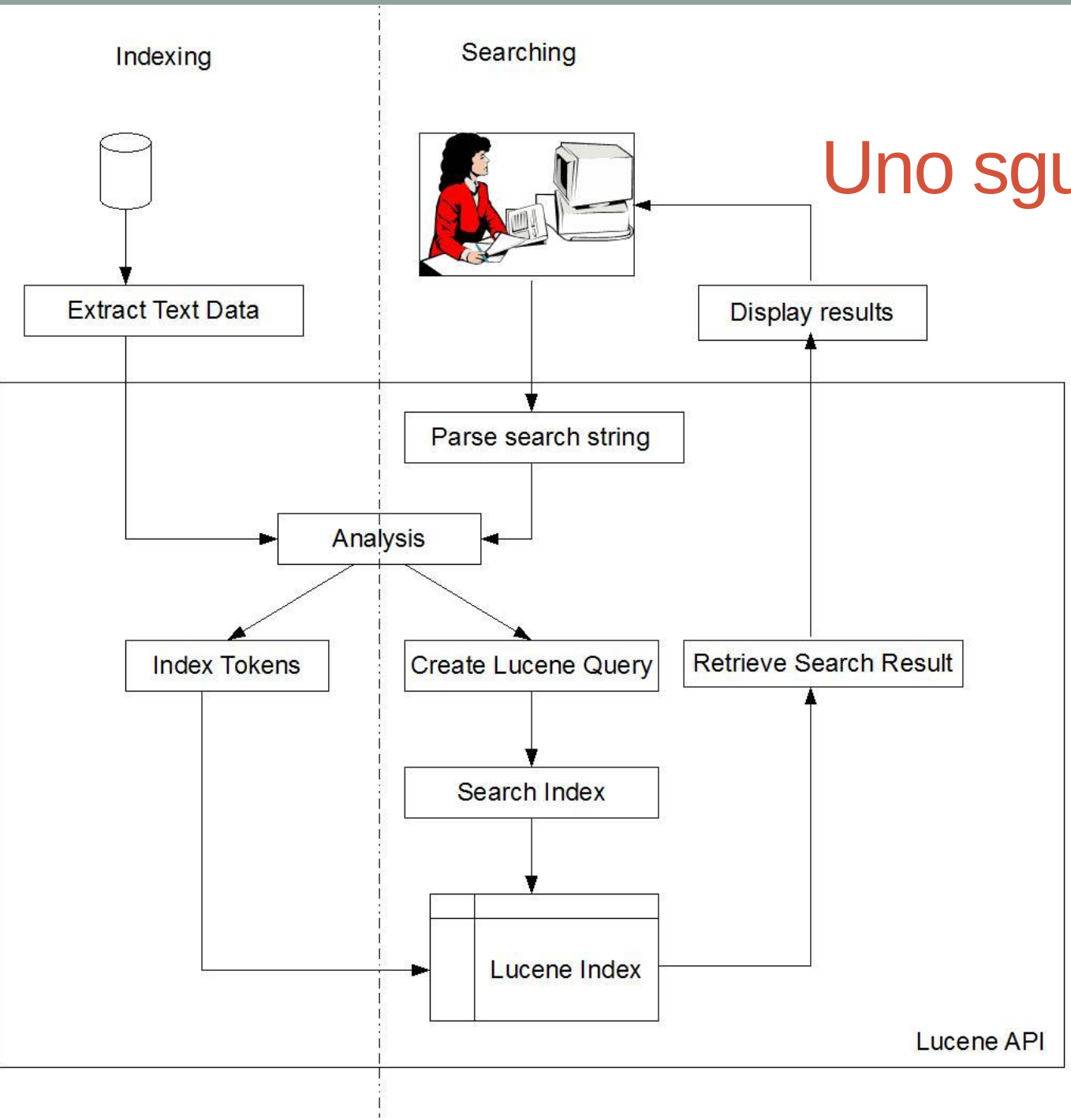
Sistemi di IR: Componenti

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

Operazioni sui Testi.



Uno sguardo da dentr



Sistemi di IR: Componenti (2)

- INDICIZZAZIONE
  - Costruisce l'indice inverso: parole
    → riferimenti ai documenti
- 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.

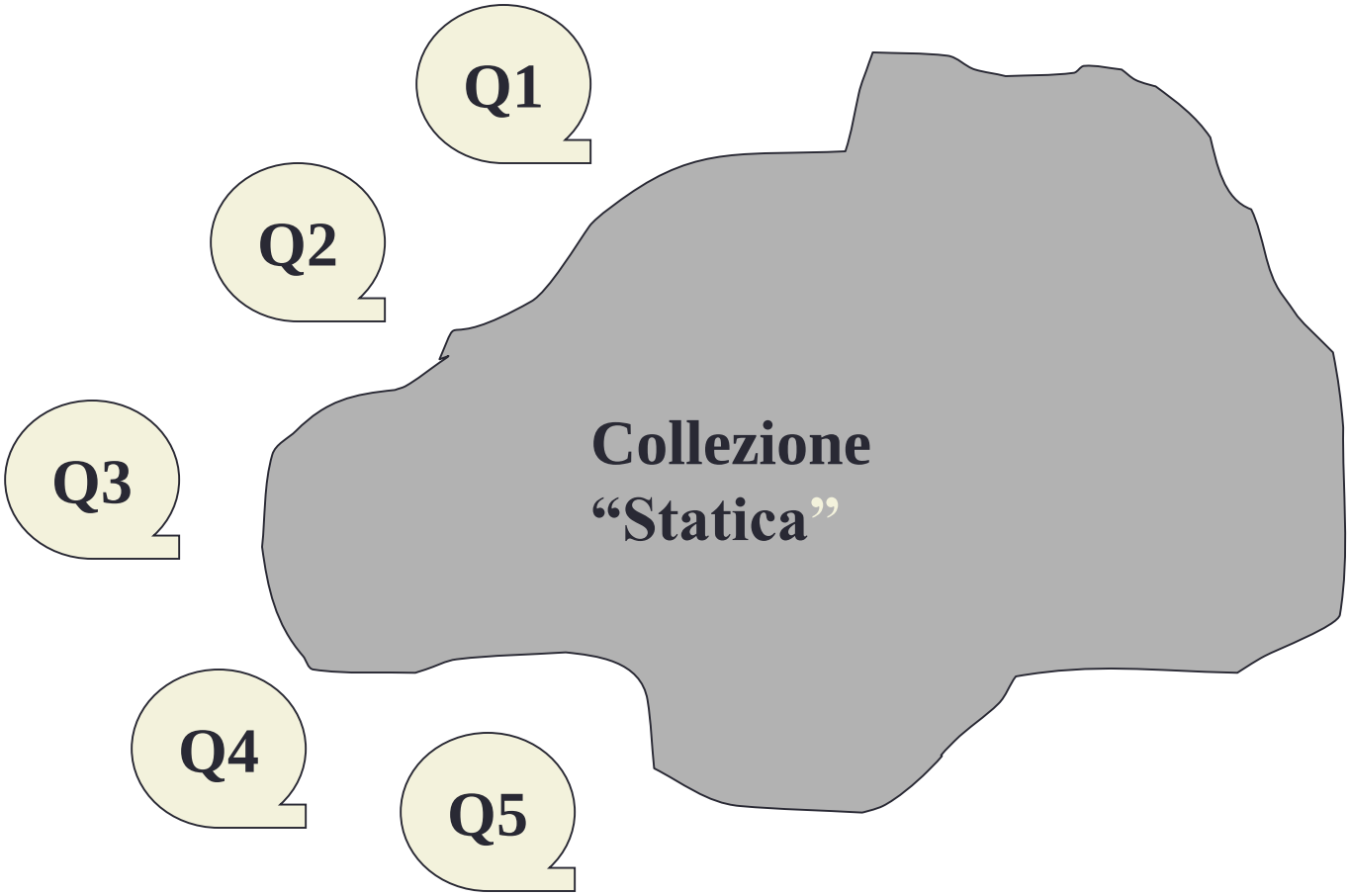
IR: Ulteriori *task*

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

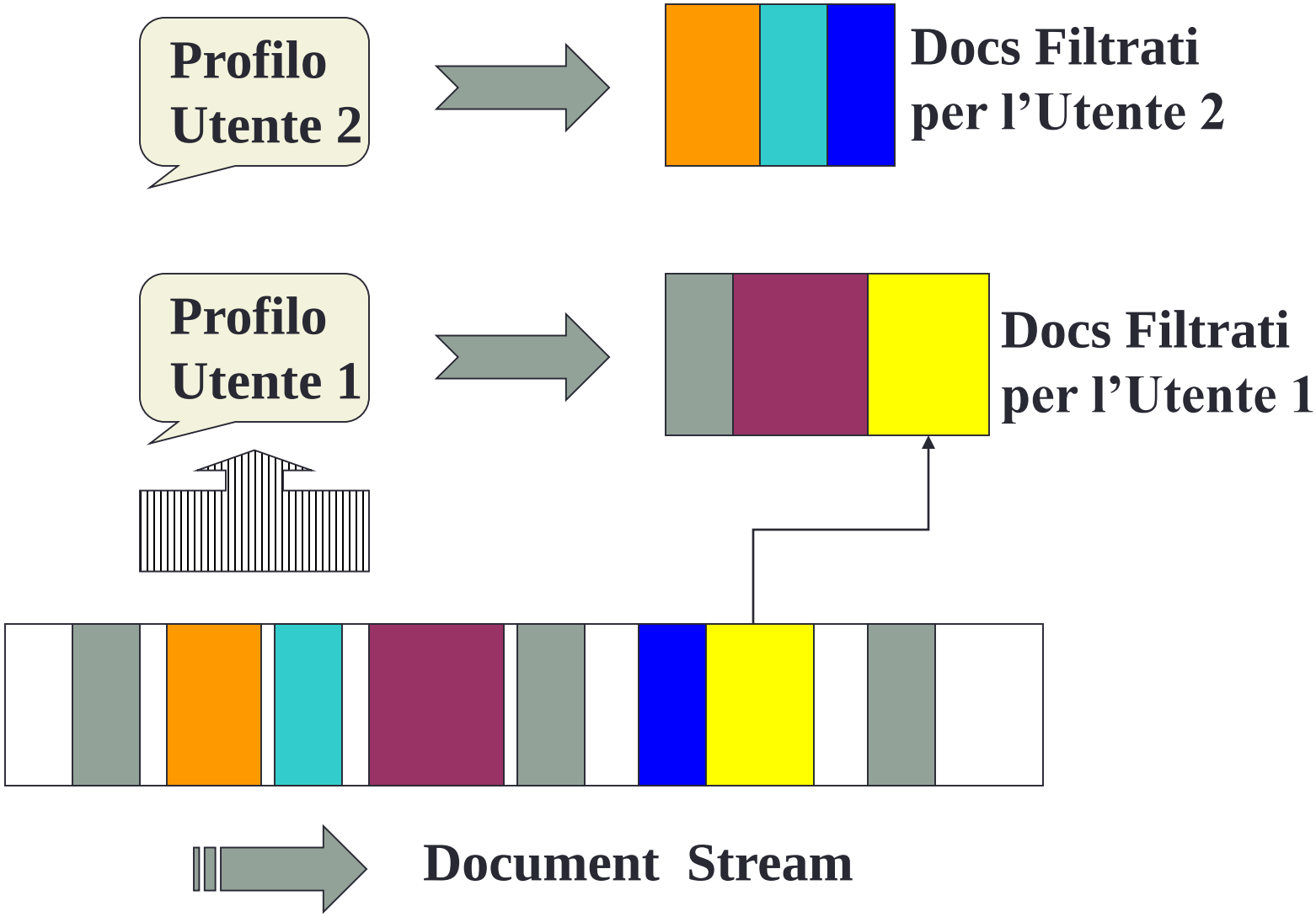
Retrieval Tasks

- *Ad hoc retrieval*: Collezione di Documenti stabile ed interrogazioni variabili.
- *Filtering*: Query (pre)fissate e flussi continui di documenti
  - Profilo Utente: un modello (statico) di preferenze relative.
  - Target: decisione binaria.
- Routing: come il *filtering* ma con funzioni di rilevanza/adesione alle preferenze non binari e generalmente dinamici.

Ad Hoc Retrieval



Filtering


Profilo Utente 2
Docs Filtrati per l'Utente 2
Profilo Utente 1
Docs Filtrati per l'Utente 1
Document Stream

IR: Contiguita' Disciplinare

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

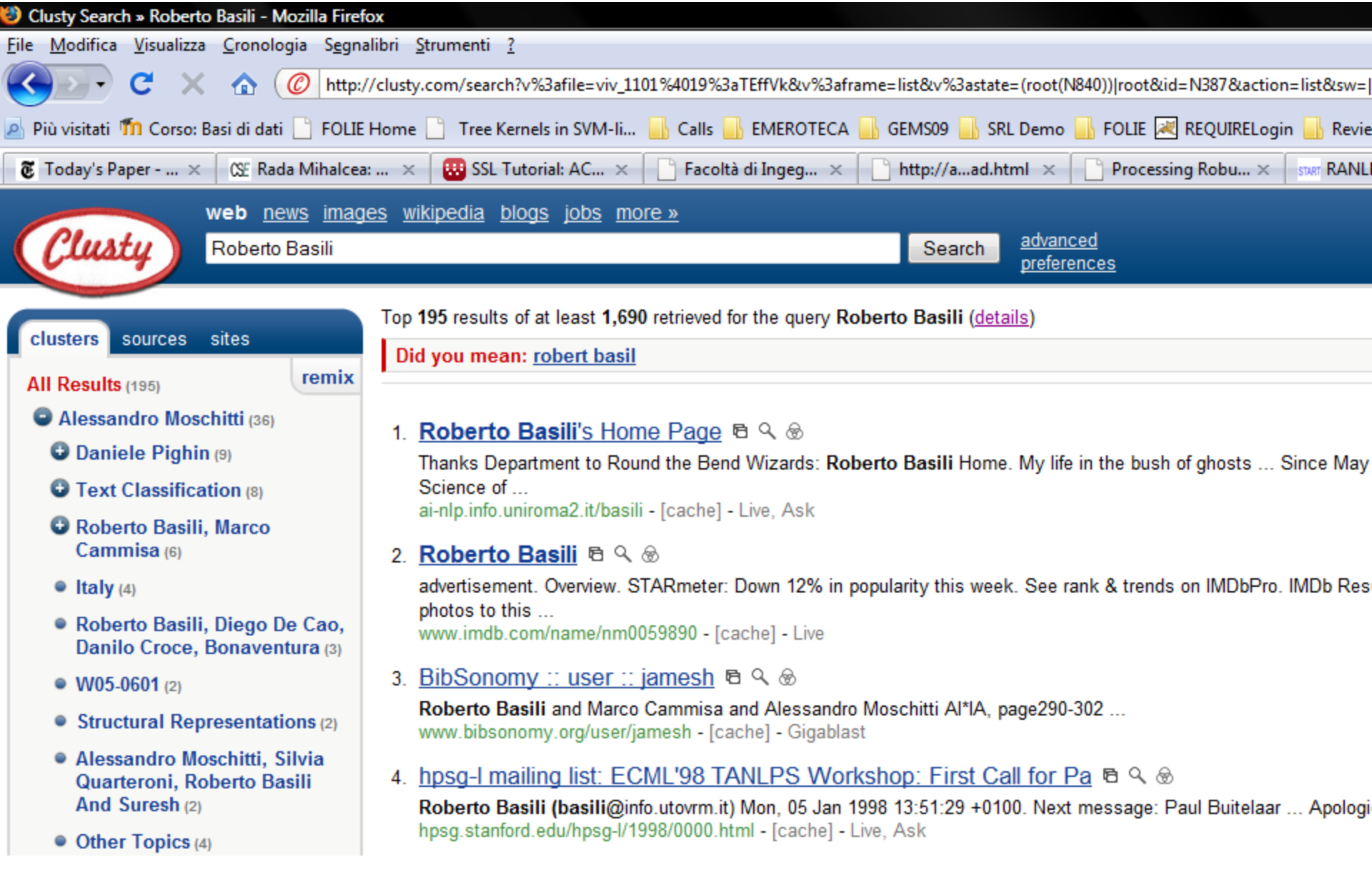
Machine Learning

- 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*).

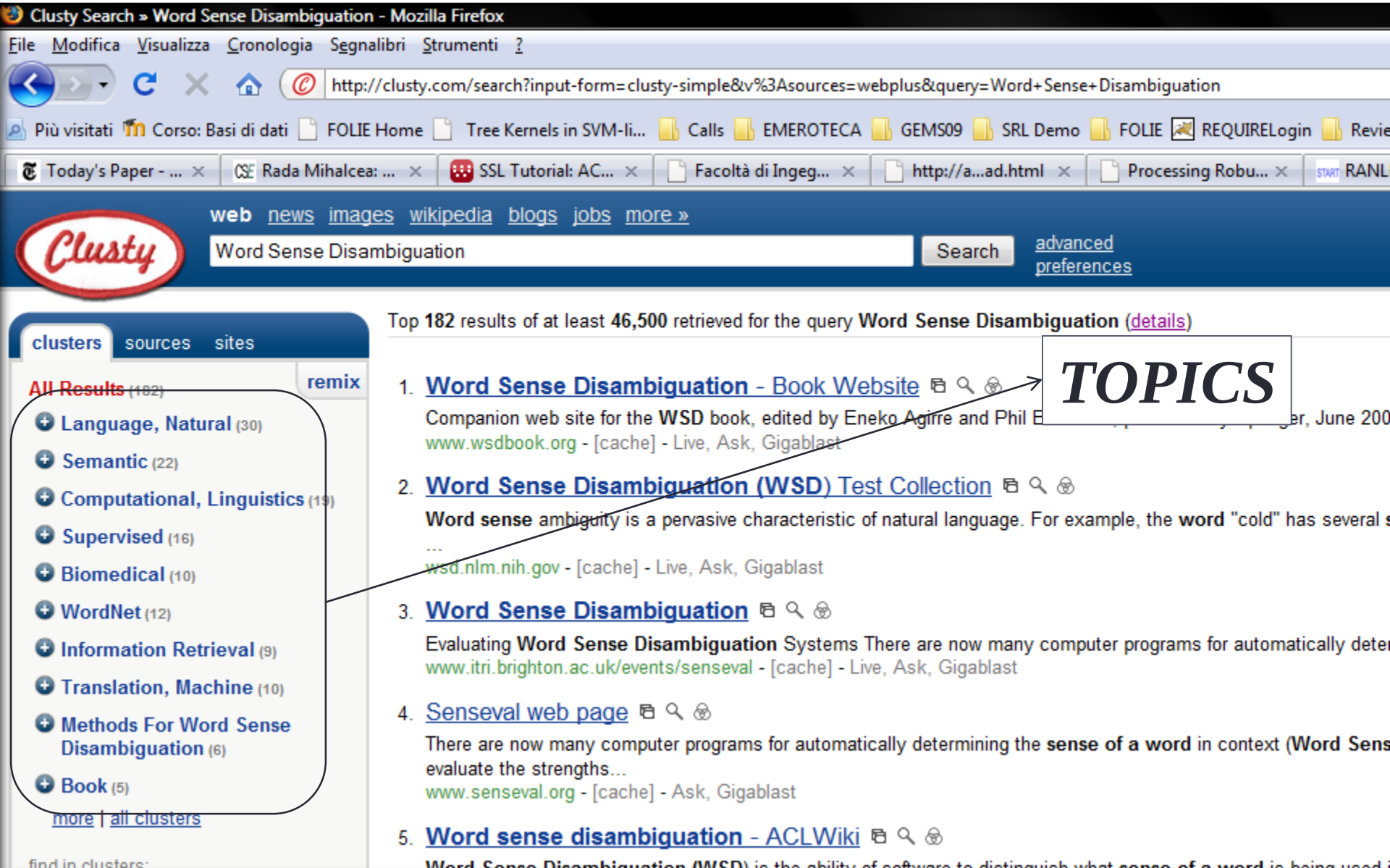
Machine Learning: direzioni verso l'IR

- **Categorizzazione** dei Testi
  - Classificazione Automatica Gerarchica (es. Yahoo).
  - Filtering/Routing/Reccomendation Adattivi
  - Automated *spam filtering*.
- **Clustering** dei Testi
  - Clustering dei risultatio 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



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
Today's Paper - ... Rada Mihalcea: ... SSL Tutorial: AC... Facoltà di Ingeg... http://a...ad.html Processing Robu... RANL
Clusty
web news images wikipedia blogs jobs more »
Word Sense Disambiguation
Search
advanced preferences
clusters sources sites
remix
All Results (182)
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)
TOPICS
1. Word Sense Disambiguation - Book Website
Companion web site for the WSD book, edited by Eneko Agirre and Phil E
www.wsdbook.org - [cache] - Live, Ask, Gigablast
2. Word Sense Disambiguation (WSD) Test Collection
Word sense ambiguity is a pervasive characteristic of natural language. For example, the word "cold" has several
...
wsd.nlm.nih.gov - [cache] - Live, Ask, Gigablast
3. Word Sense Disambiguation
Evaluating Word Sense Disambiguation Systems There are now many computer programs for automatically deter
www.itri.brighton.ac.uk/events/senseval - [cache] - Live, Ask, Gigablast
4. Senseval web page
There are now many computer programs for automatically determining the sense of a word in context (Word Sens
evaluate the strengths...
www.senseval.org - [cache] - Ask, Gigablast
5. Word sense disambiguation - ACLWiki

Sommario

- Perche' l'IR è importante
- Cos'e' 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