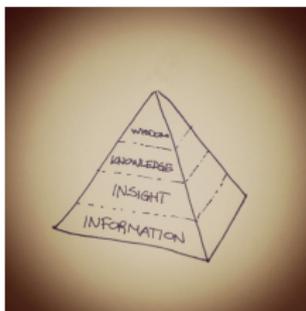


# Information retrieval e motori di ricerca ( in riferimento a [SA15, Cap. 4] )

EUGENIO OMODEO  
Università degli Studi di Trieste.



Trieste, 4.11.2020

# Sunto

Cominciamo con la discussione dei sistemi d' *information retrieval* prima di passare ai **motori di ricerca**.<sup>1</sup>

---

<sup>1</sup>Per i secondi, una comparazione tecnica è piuttosto difficile dato che ogni motore realizza le proprie funzionalità mediante tecniche proprietarie i cui dettagli non vengono resi pubblici.

## Visione di un importante inventore del ?? sec.

*Much needs to occur, however, between the collection of data and observations, the extraction of parallel material from the existing record, and the final insertion of new material into the general body of the common record.*

*For mature thought there is no mechanical substitute.*

*But creative thought and essentially repetitive thought are very different things. For the latter there are, and may be, powerful mechanical aids.*

[...]

## Visione di un importante inventore del XX sec.

*We seem to be worse off than before —  
for we can enormously extend the record; yet even in its  
present bulk we can hardly consult it.*

*This is a much larger matter than merely the extraction  
of data for the purposes of scientific research; it involves  
the entire process by which man profits by his inheritance  
of acquired knowledge. The prime action of use is  
selection, and here we are halting indeed. There may be  
millions of fine thoughts, and the account of the  
experience on which they are based, all encased within  
stone walls of acceptable architectural form; but if the  
scholar can get at only one a week by diligent search, his  
syntheses are not likely to keep up with the current scene.*

[...]

# Visione di un importante inventore del XX sec.

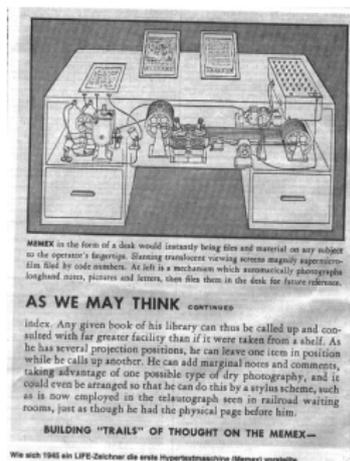
*Selection, in this broad sense, is a stone adze in the hands of a cabinetmaker.*

Vannevar Bush, *As we may think*, 1945



**La selezione, in questo senso lato, è un'ascia di pietra in mano a un ebanista.**

# Il MEMEX di Vannevar Bush



**Vannevar Bush** ( Everett, 11 mar 1890 — Belmont, 30 giu 1974 ) è stato uno scienziato e tecnologo statunitense. Fu un inventore e coordinò le attività di ricerca degli USA durante la seconda guerra mondiale; precursore degli ipertesti, è stato l'ideologo del supporto delle attività di ricerca ai fini del potenziamento delle democrazie.

# Vannevar Bush

( 1890–1974 )



# Scaletta

## *Information retrieval*

Mansioni

Organizzazione interna

## *Tecniche d'information retrieval*

Tecniche di analisi

Tecniche d'indicizzazione e di *matching*

## Cos'è l'Information retrieval ?

*I sistemi d'Information Retrieval sono stati concepiti con l'obiettivo di mediare l'interazione fra l'utente e il corpus di documenti che egli desidera interrogare.*

*Tipicamente, l'utente sottopone al sistema una o più chiavi di ricerca ( keyword ) che denotano il suo bisogno d'informazione e il sistema, consultando il corpus, restituisce l'insieme di documenti che sono valutati come pertinenti rispetto alla richiesta.*

[CFM09, pag. 66]

# Efficacia ed efficienza nell'*Information retrieval*

- ▶ L'utente desidera una risposta accurata alla propria richiesta, cioè costituita da *tutti* i documenti del corpus che sono rilevanti.
- ▶ Desidera anche un *basso tempo d'attesa* tra la formulazione della richiesta e la ricezione del risultato.

Questi obiettivi sono tra loro contrastanti

## Rappresentazione di sintesi nell'*Information retrieval*

I sistemi d'I.R. eseguono operazioni di manipolazione del corpus per estrarne una **rappresentazione di sintesi** del contenuto informativo di ciascun documento e memorizzare tali rappresentazioni in modo efficiente.

L'utente non ha accesso diretto alle rappres. di sintesi

# Esempio d'interrogazione ad un sistema d'I.R.

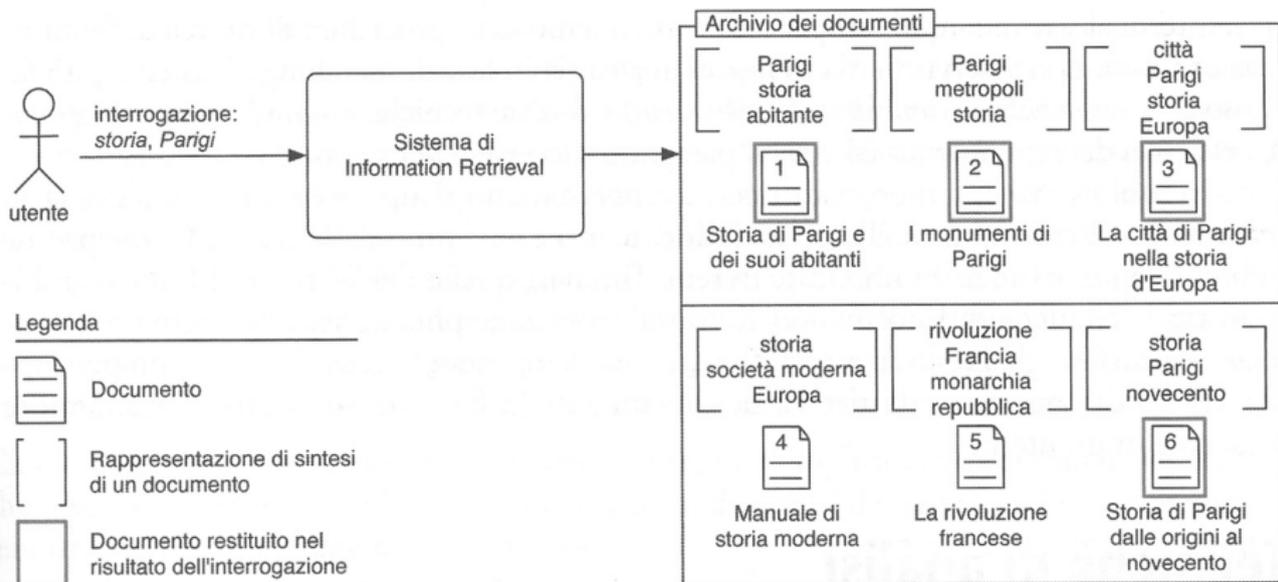
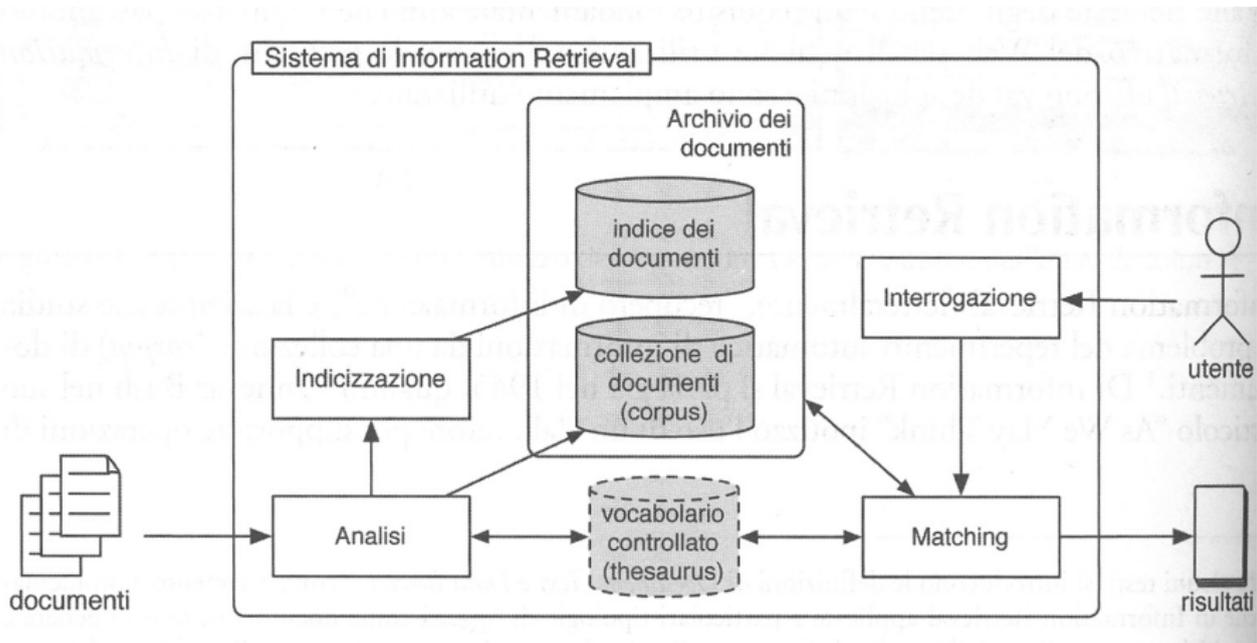


Figura 7.2 Esempio di Information Retrieval.

# Articolazione di un sistema d'Information retrieval



# Analisi e indicizzazione

Le componenti di **analisi e indicizzazione**

- ▶ vengono invocate allorché nuovi documenti vengono aggiunti al sistema e
- ▶ alimentano l'**archivio dei documenti**.

**Analisi:** Elabora il documento originale memorizzandolo nel **corpus** e producendone una rappresentazione di sintesi.

**Indicizzazione:** Ricava dalla rappresentazione di sintesi un **indice**, i.e. una struttura di accesso efficiente ai documenti.

# Corpus e indice

Almeno a livello concettuale, i due oggetti sono ben diversi:

**Corpus:** è l'insieme dei documenti  
inseriti nel sistema;

**Indice:** è una struttura basata sulle  
rappresentazioni di sintesi.

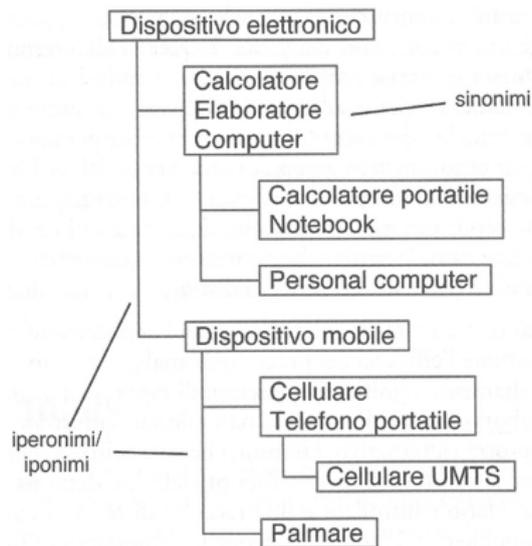
# Interrogazione e *matching* 🗉 “far combaciare”

**Interrogazione:** È la componente cui l'utente somministra le proprie richieste: essa le prepara per la ricerca.

**Matching:** È la componente che confronta ogni richiesta con l'indice, per individuare nel corpus i documenti rilevanti: il risultato.

# Thesaurus

Le fasi di analisi e di *matching* potranno far uso di un *thesaurus* e di altre tecniche linguistiche per migliorare l'efficacia del reperimento.



di vocabolario controllato.

# Un'applicaz. di tecniche d'I.R. ( Si pensi pure a Spotlight )



## Inside Google Desktop

The official source for information about Google Desktop.

## Google Desktop Update

Friday, September 02, 2011 12:48 PM

In 2004, Google launched Google Desktop, a program designed to make it easy for users to search their own PCs for emails, files, music, photos, Web pages and more.

Desktop has been used by tens of millions of people and we've been humbled by its usage and great user feedback. However, over the past seven years we've also witnessed some big changes in how users store and access their own data, with many moving to web-based applications. There has been a significant shift from local to cloud-based storage and computing, as well as integration of Google Desktop functionality (like local search) into most modern operating systems. This is a positive development for users and we're excited that most people now have instant access to their personal information. As such, we'll be discontinuing support for Google Desktop, including all of the associated APIs, services, plugins and gadgets.

## Eliminazione delle *stop word*

Si tratta di particelle ( articoli, preposiz., congiunz. ) la cui eliminazione non ha importanti ripercussioni sul contenuto informativo, ma può ridurre significativamente ( anche dimezzandola ) la lunghezza del testo.

Eliminazione delle *stop word*

## Esempio dell'italiano

<http://snowball.tartarus.org/algorithms/italian/stop.txt>

| An Italian stop word list. Comments begin with vertical bar. Each stop  
| word is at the start of a line.

```
ad          | a (to) before vowel
al          | a + il
allo       | a + lo
ai         | a + i
agli       | a + gli
all        | a + l'
agl        | a + gl'
alla       | a + la
alle       | a + le
con        | with
col        | con + il
coi        | con + i (forms collo, cogli etc are now very rare)
da         | from
```

## Estrazione di *stem*

Il dizionario italiano HOEPLI definisce **tema** e **lemma**,<sup>2</sup> nel senso della morfologia, rispettivam. così:

- 5 LING Parte fissa della parola, alla quale si salda la desinenza producendo la flessione
- 3 LING Ognuna delle voci definite da un dizionario o da un'enciclopedia

Il processo di *stemming* consiste nel sostituire tutte le forme derivate con il tema corrispondente, per favorire l'estrazione di parole ad elevata rappresentatività.

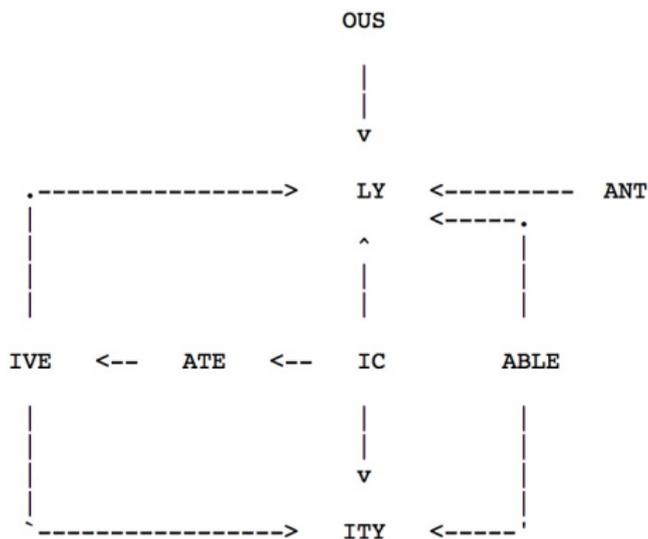
---

<sup>2</sup>Spesso useremo 'termine' per 'lemma'.

# Combinazione di desinenze

# Esempio

<http://snowball.tartarus.org/texts/romance.html>



## Abolizione di desinenze

## Esempi

<http://snowball.tartarus.org/texts/romance.html>

In English, ABLE combines with LY to form ABLY. So in French, for example, *able* combines with (*e*)*ment* to form *ablement*. In some languages particular combinations are rare. In Italian, for example, ANT + LY, which would be the ending *antemente*, is so rare that it does not figure in the stemming algorithm. According to the picture, we should encounter the forms ICATIVELY and ICATIVITY, and dictionaries instance a few English words with these endings (*communicatively* for example). But in practice three is the maximum number of derivational suffixes that one need consider in combination.

Estrazione di *stem*

## Esempio lingue romanze

<http://snowball.tartarus.org/texts/romance.html>

The *d*-suffixes of all four languages follow a similar pattern.  
They can be tabulated as follows,

		French	Spanish	Portug.	Italian
noun	ANCE	<i>ance</i>	<i>anza</i>	<i>eza</i>	<i>anza</i>
adjective	IC	<i>ique</i>	<i>ico</i>	<i>ico</i>	<i>ico</i>
noun	ISM	<i>isme</i>	<i>ismo</i>	<i>ismo</i>	<i>ismo</i>
adjective	ABLE	<i>able</i>	<i>able</i>	<i>ável</i>	<i>abile</i>
adjective	IBLE	-	<i>ible</i>	<i>ível</i>	<i>ibile</i>
noun	IST	<i>iste</i>	<i>ista</i>	<i>ista</i>	<i>ista</i>
adjective	OUS	<i>eux</i>	<i>oso</i>	<i>oso</i>	<i>oso</i>
noun	MENT	<i>ment</i>	<i>amiento</i>	<i>amento</i>	<i>mente</i>
noun	ATOR	<i>ateur</i>	<i>ador</i>	<i>ador</i>	<i>attore</i>
noun	ATRESS	<i>atrice</i>	-	-	<i>atrice</i>
noun	ATION	<i>ation</i>	<i>ación</i>	<i>ação</i>	<i>azione</i>

## Scelta di termini ad elevato potere discriminante

Queste tecniche mirano a estrarre i termini che meglio rappresentano il contenuto informativo di un documento. Se il corpus è

- eterogeneo**, i.e. costituito da documenti riguardanti argomenti vari, verranno selezionati come **significativi**, all'interno di ciascun documento, quei termini che vi occorrono con maggiore frequenza.
- omogeneo**: verranno selezionati come **distintivi** di ciascun documento quei termini che ricorrono frequentemente in un documento ma raramente nel corpus.

In alternativa...

... ci si rifarà a un **thesaurus**

Un **thesaurus** correla termini mediante relazioni di

- ▶ **sinonimía**, come ad es. ( pressappoco ) **morsel** / **mouthful**
- ▶ **iperonimía** / **iponimía**, come ad es. **computer** / **elaboratore**
- ▶ **meronimía** / **olonimía**, come ad es. **dito** / **mano**, **ruota** / **auto**
- ▶ ecc.

## Formulazione manuale / automatica di un thesaurus

Utilizzando strumenti quali <http://wordnet.princeton.edu>, un utente esperto potrà associare manualmente a ogni documento una lista di termini che ne rappresentano il contenuto informativo.

In alternativa, ci si potrà avvalere di strumenti automatici di natura statistica.

In un approccio combinato, si procederà dai risultati automatici a una convalida / revisione manuale.



# Indicizzazione

L'**indice** di un corpus è costituito da coppie:

$(l_i, R_i)$  dove ogni  $l_i$  è un termine e  
il corrispondente  $R_i$  riferisce l'insieme dei documenti  
collegati a  $l_i$ .

Certe tecniche memorizzano negli  $R_i$  oltre ai documenti riferiti  
anche:

- ▶ la *frequenza* con cui  $l_i$  *occorre* all'interno di  $R_i$ ;
- ▶ in *quali parti* di ciascun documento figura  $l_i$ ;
- ▶ la *vicinanza* nello stesso documento di altri termini.

# Matching

L'interfaccia d'**interrogazione** permette all'utente finale di formulare una lista di **chiavi di ricerca** ( usualmente in congiunzione ).

Perché un documento sia restituito nel risultato, il sistema deve trovare almeno un termine che *combaci* con ciascuna chiave.

Tramite tecniche preparatorie dette di **normalizzazione** verrà costruita una lista di termini di cui effettuare la ricerca.

## Matching esatto o per similarità ?

La ricerca di *matching esatto* è piú semplice, ma in genere meno soddisfacente; quella per similarità può basarsi su tecniche

*sintattiche*, quali la *distanza di editing* che fa apparire **cittadella** distante **3** da **cittadina**, perché basta cambiare tre caratteri per ottenere una dall'altra.

*linguistiche*, che tengono conto di eventuali relazioni terminologiche fra quanto cercato e quanto esaminato. Ad es:

- ▶ **città** e **centro urbano** hanno somiglianza massima, in quanto sinonimi;
- ▶ **metropoli** e **centro urbano** sono molto vicini, in quanto correlati da iper-/ipo-nimia.

# Il risultato di un'interrogazione

Il risultato sarà composto da una lista di documenti con un valore di **rilevanza** associato a ciascuna voce della lista.

	Documenti rilevanti	Documenti non rilevanti
Documenti reperiti	A – documenti reperiti e rilevanti	B – documenti reperiti ma non rilevanti (falsi positivi)
Documenti non reperiti	C – documenti non reperiti, ma rilevanti (falsi negativi)	D – documenti non reperiti e non rilevanti

**Figura 7.5** Classificazione di un corpus di documenti rispetto a un'interrogazione  $I$  e a un sistema di Information Retrieval  $IR$ .

## Riferimenti bibliografici

-  Silvana Castano, Alfio Ferrara, and Stefano Montanelli.  
*Informazione, conoscenza e web — per le scienze umanistiche.*  
Pearson / Addison Wesley, 2009.
-  Lawrence Snyder and Alessandro Amoroso.  
*FLUENCY –Conoscere e usare l'informatica.*  
Pearson Italia, Milano-Torino, 5<sup>a</sup> edition, 2015.