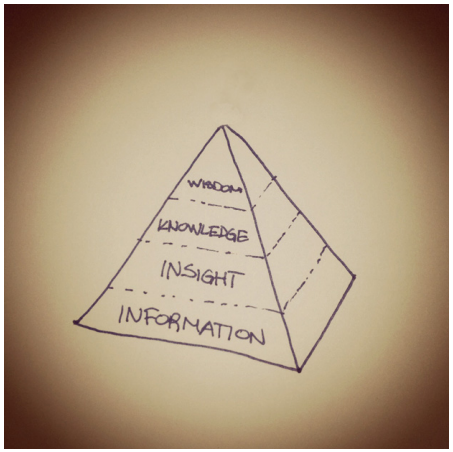
Information retrieval e motori di ricerca (in rif. a [SHA20, Cap. 4])

Eugenio Omodeo
Università degli Studi di Trieste.



Trieste, 22.11.22

Sunto

Cominciamo con la discussione dei sistemi d'*information retrieval* prima di passare ai motori di ricerca.[1]

[1] Per i secondi, una comparazione tecnica è piuttosoto 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.

$[\cdots]$

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.*

$[\cdots]$

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

A SCIENTIST OF THE FUTURE RECORDS EXPERIMENTS WITH A TINY CAMERA FITTED WITH UNIVERSAL-FOCUS LENS. THE SMALL SQUARE IN THE EYEGLASS AT THE LEFT SIGHTS THE OBJ

AS WE MAY THINK

A TOP U. S. SCIENTIST FORESEES A POSSIBLE FUTURE WORLD IN WHICH MAN-MADE MACHINES WILL START TO THINK

by VANNEVAR BUSH

DIRECTOR OF THE OFFICE OF SCIENTIFIC RESEARCH AND DEVELOPMENT

Condensed from the *Atlantic Monthly*, July 1945

This has not been a scientists' war; it has been a war in which all have had a part. The scientists, burying their old professional competition in the demand of a common cause, have shared greatly and learned much. It has been exhilarating to work in effective partnership. What are the scientists to do next?

For the biologists, and particularly for the medical scientists, there can be little indecision, for their war work has hardly required them to leave the old paths. Many indeed have been able to carry on their war research in their familiar peacetime laboratories. Their objectives remain much the same.

It is the physicists who have been thrown most violently off stride, who have left academic pursuits for the making of strange destructive gadgets, who have had to devise new methods for their unanticipated assignments. They have done their part on the devices that made it possible to turn back the enemy. They have worked in combined effort with the physicists of our allies. They have felt within themselves the stir of achievement. They have been part of a great team. Now one asks whest they will find objectives worthy of their best.

* * *

There is a growing mountain of research. But there is increased evidence that we are being bogged down today as specialization extends. The investigator is staggered by the findings and conclusions of thousands of other workers—conclusions which he cannot find time to grasp, much less to remember, as they appear. Yet specialization becomes increasingly necessary for progress, and the effort to bridge between disciplines is correspondingly suspicial.

Professionally our methods of transmitting and reviewing the results research are generations old and by now are totally inadequate for their pose. If the aggregate time spent in writing scholarly works and in read them could be evaluated, the ratio between these amounts of time might well be startling. Those who conscientiously attempt to keep abreast of rent thoughts, even in restricted fields, by close and continuous reading might well shy away from an examination calculated to show how much of the vious month's efforts could be produced on call.

Mendel's concept of the laws of genetics was lost to the world for a eration because his publication did not reach the few who were capof grasping and extending it. This sort of catastrophe is undoubtedly repeated all about us as truly significant attainments become lost in the of the inconsequential.

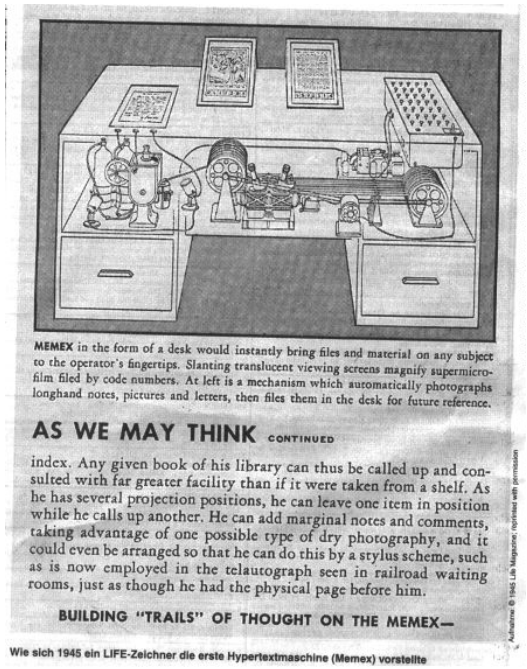
Publication has been extended far beyond our present ability to make use of the record. The summation of human experience is being expanded a prodigious rate, and the means we use for threading through the quent maze to the momentarily important item is the same as was used the days of square-rigged ships.

But there are signs of a change as new and powerful instrumentalities come into use. Photocells capable of seeing things in a physical sense, vanced photography which can record what is seen or even what is thermionic tubes capable of controlling potent forces under the guidance

112

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

Il MEMEX di Vannevar Bush



MEMEX in the form of a desk would instantly bring files and material on any subject to the operator's fingertips. Slanting translucent viewing screens magnify supermicrofilm filed by code numbers. At left is a mechanism which automatically photographs longhand notes, pictures and letters, then files them in the desk for future reference.

AS WE MAY THINK CONTINUED

index. Any given book of his library can thus be called up and consulted with far greater facility than if it were taken from a shelf. As he has several projection positions, he can leave one item in position while he calls up another. He can add marginal notes and comments, taking advantage of one possible type of dry photography, and it could even be arranged so that he can do this by a stylus scheme, such as is now employed in the telautograph seen in railroad waiting rooms, just as though he had the physical page before him.

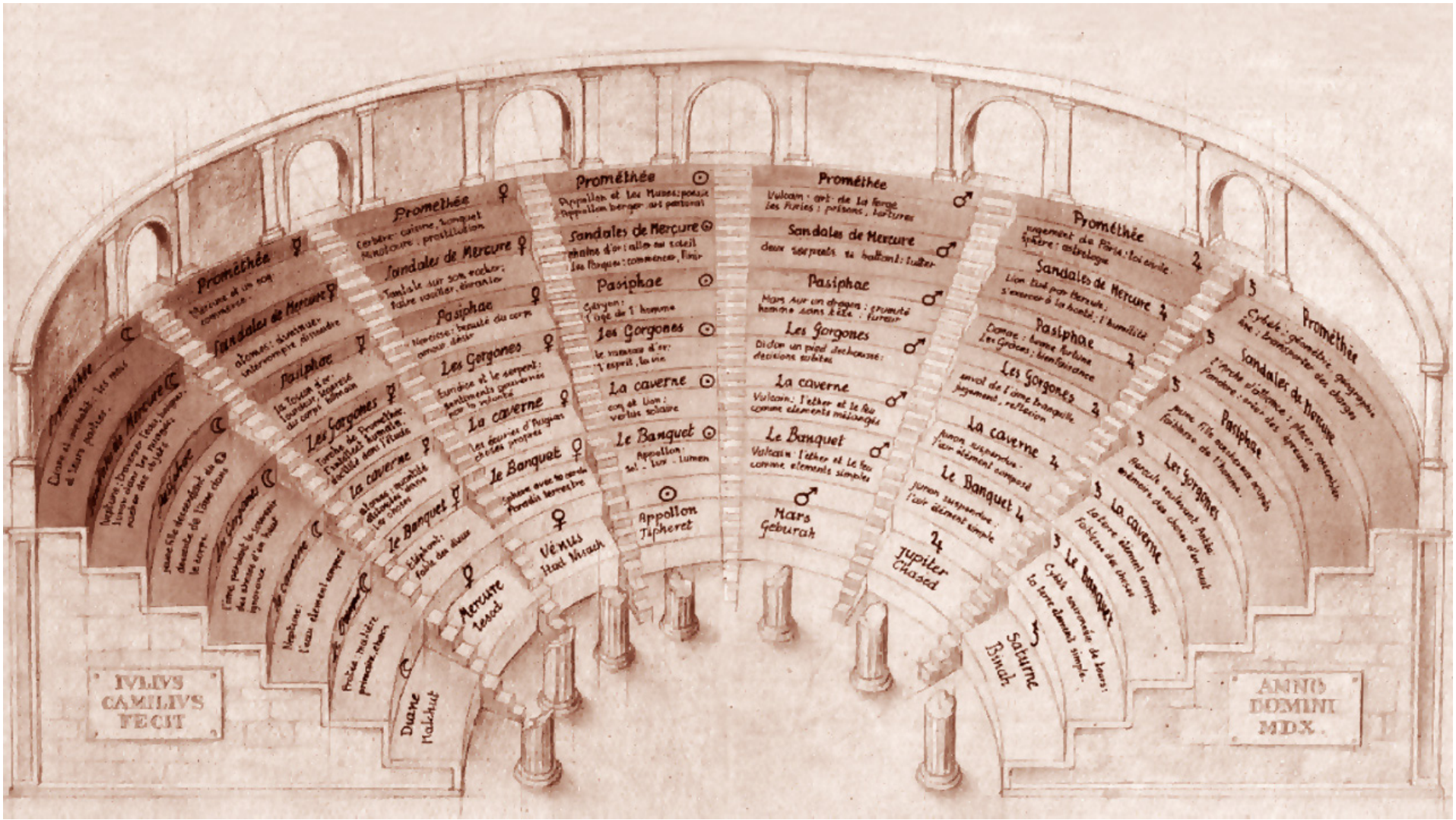
BUILDING "TRAILS" OF THOUGHT ON THE MEMEX—

Wie sich 1945 ein LIFE-Zeichner die erste Hypertextmaschine (Memex) vorstellte

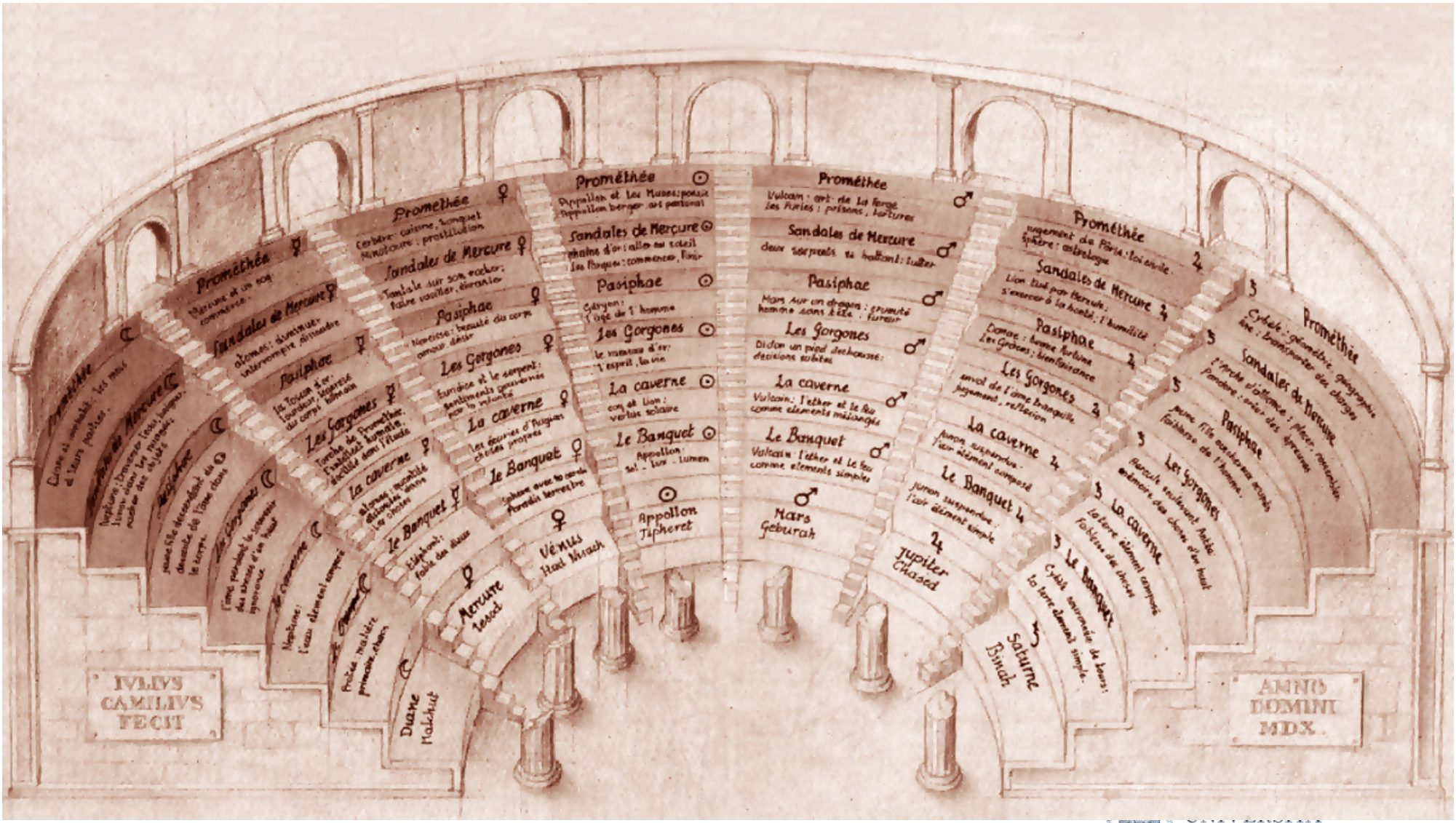
Vannevar Bush (Everett, 11 mar 1890 — Belmont, 30 giu 1974) è stato uno scienziato e tecnologo statunitense. Fu un inventore e coordinò le attività di ricerca degli USA durante la seconda guerra mondiale; precursore degli ipertesti, è stato l'ideologo del supporto delle attività di ricerca ai fini del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

Vannevar Bush (1890–1974)

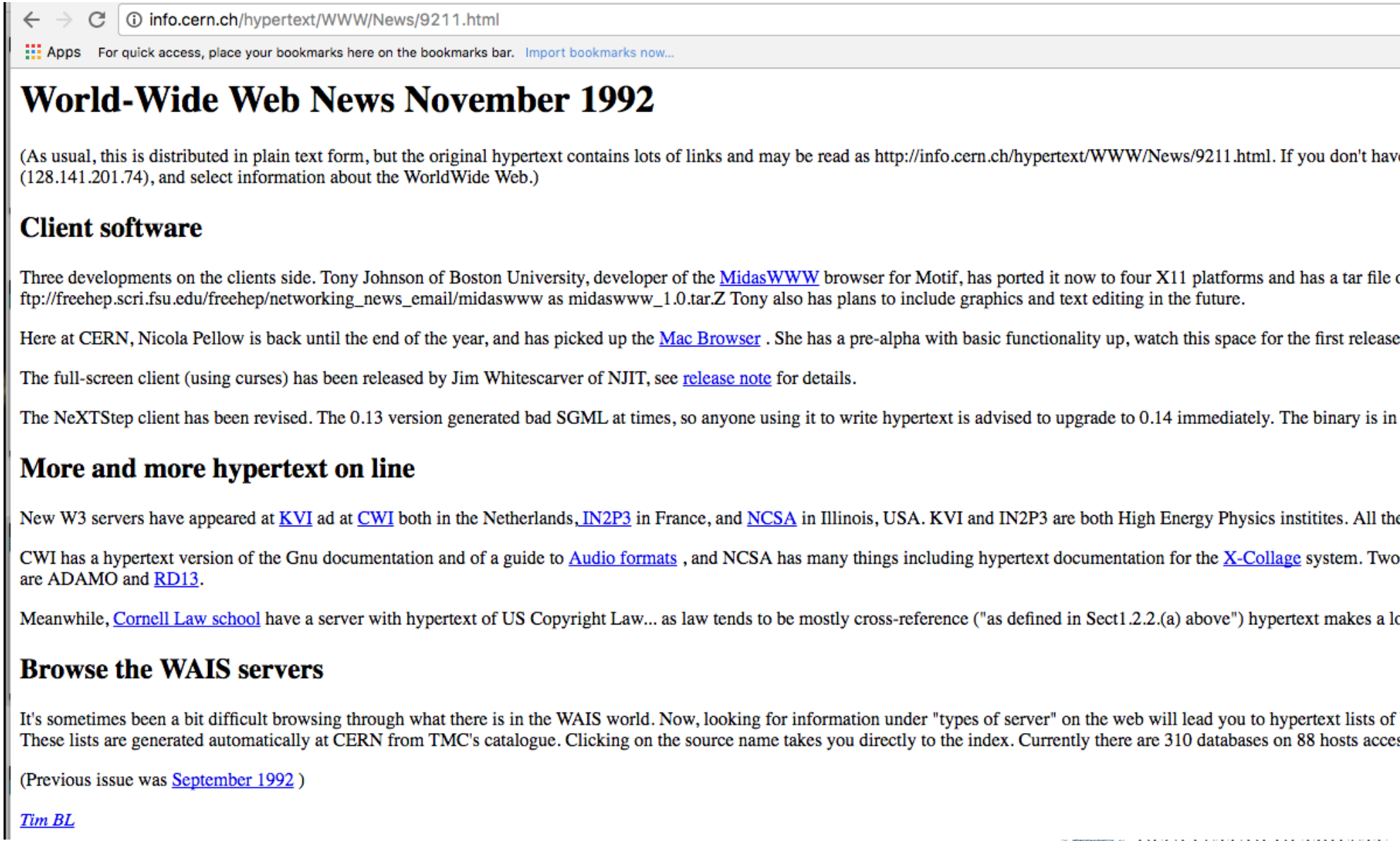
Un precursore del XVI secolo: Giulio Camillo Delminio



Teatro della memoria del XVI secolo



WWW 1992



info.cern.ch/hypertext/WWW/News/9211.html

Apps For quick access, place your bookmarks here on the bookmarks bar. Import bookmarks now...

World-Wide Web News November 1992

(As usual, this is distributed in plain text form, but the original hypertext contains lots of links and may be read as http://info.cern.ch/hypertext/WWW/News/9211.html. If you don't hav
(128.141.201.74), and select information about the WorldWide Web.)

Client software

Three developments on the clients side. Tony Johnson of Boston University, developer of the MidasWWW browser for Motif, has ported it now to four X11 platforms and has a tar file
ftp://freehep.scri.fsu.edu/freehep/networking_news_email/midaswww as midaswww_1.0.tar.Z Tony also has plans to include graphics and text editing in the future.

Here at CERN, Nicola Pellow is back until the end of the year, and has picked up the Mac Browser . She has a pre-alpha with basic functionality up, watch this space for the first release

The full-screen client (using curses) has been released by Jim Whitescarver of NJIT, see release note for details.

The NeXTStep client has been revised. The 0.13 version generated bad SGML at times, so anyone using it to write hypertext is advised to upgrade to 0.14 immediately. The binary is in

More and more hypertext on line

New W3 servers have appeared at KVI ad at CWI both in the Netherlands, IN2P3 in France, and NCSA in Illinois, USA. KVI and IN2P3 are both High Energy Physics institites. All the

CWI has a hypertext version of the Gnu documentation and of a guide to Audio formats , and NCSA has many things including hypertext documentation for the X-Collage system. Two
are ADAMO and RD13.

Meanwhile, Cornell Law school have a server with hypertext of US Copyright Law... as law tends to be mostly cross-reference ("as defined in Sect1.2.2.(a) above") hypertext makes a lo

Browse the WAIS servers

It's sometimes been a bit difficult browsing through what there is in the WAIS world. Now, looking for information under "types of server" on the web will lead you to hypertext lists of
These lists are generated automatically at CERN from TMC's catalogue. Clicking on the source name takes you directly to the index. Currently there are 310 databases on 88 hosts acces

(Previous issue was September 1992)

Tim BL

DEGLI STUDI DI TRIESTE

WWW 1992

info.cern.ch/hypertext/WWW/News/9211.html

Apps For quick access, place your bookmarks here on the bookmarks bar. Import bookmarks now...

World-Wide Web News November 1992

(As usual, this is distributed in plain text form, but the original hypertext contains lots of links and may be read as http://info.cern.ch/hypertext/WWW/News/9211.html. If you don't hav
(128.141.201.74), and select information about the WorldWide Web.)

Client software

Three developments on the clients side. Tony Johnson of Boston University, developer of the MidasWWW browser for Motif, has ported it now to four X11 platforms and has a tar file
ftp://freehep.scri.fsu.edu/freehep/networking_news_email/midaswww as midaswww_1.0.tar.Z Tony also has plans to include graphics and text editing in the future.

Here at CERN, Nicola Pellow is back until the end of the year, and has picked up the Mac Browser . She has a pre-alpha with basic functionality up, watch this space for the first release

The full-screen client (using curses) has been released by Jim Whitescarver of NJIT, see release note for details.

The NeXTStep client has been revised. The 0.13 version generated bad SGML at times, so anyo binary is in

More and more hypertext on line

New W3 servers have appeared at KVI ad at CWI both in the Netherlands, IN2P3 in France, and tites. All the

CWI has a hypertext version of the Gnu documentation and of a guide to Audio formats , and N system. Two
are ADAMO and RD13.

Meanwhile, Cornell Law school have a server with hypertext of US Copyright Law... as law ten t makes a lo

Browse the WAIS servers

It's sometimes been a bit difficult browsing through what there is in the WAIS world. Now, look text lists of
These lists are generated automatically at CERN from TMC's catalogue. Clicking on the source 3 hosts acces

(Previous issue was September 1992)

Tim BL

DEGLI STUDI DI TRIESTE

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 piú 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.

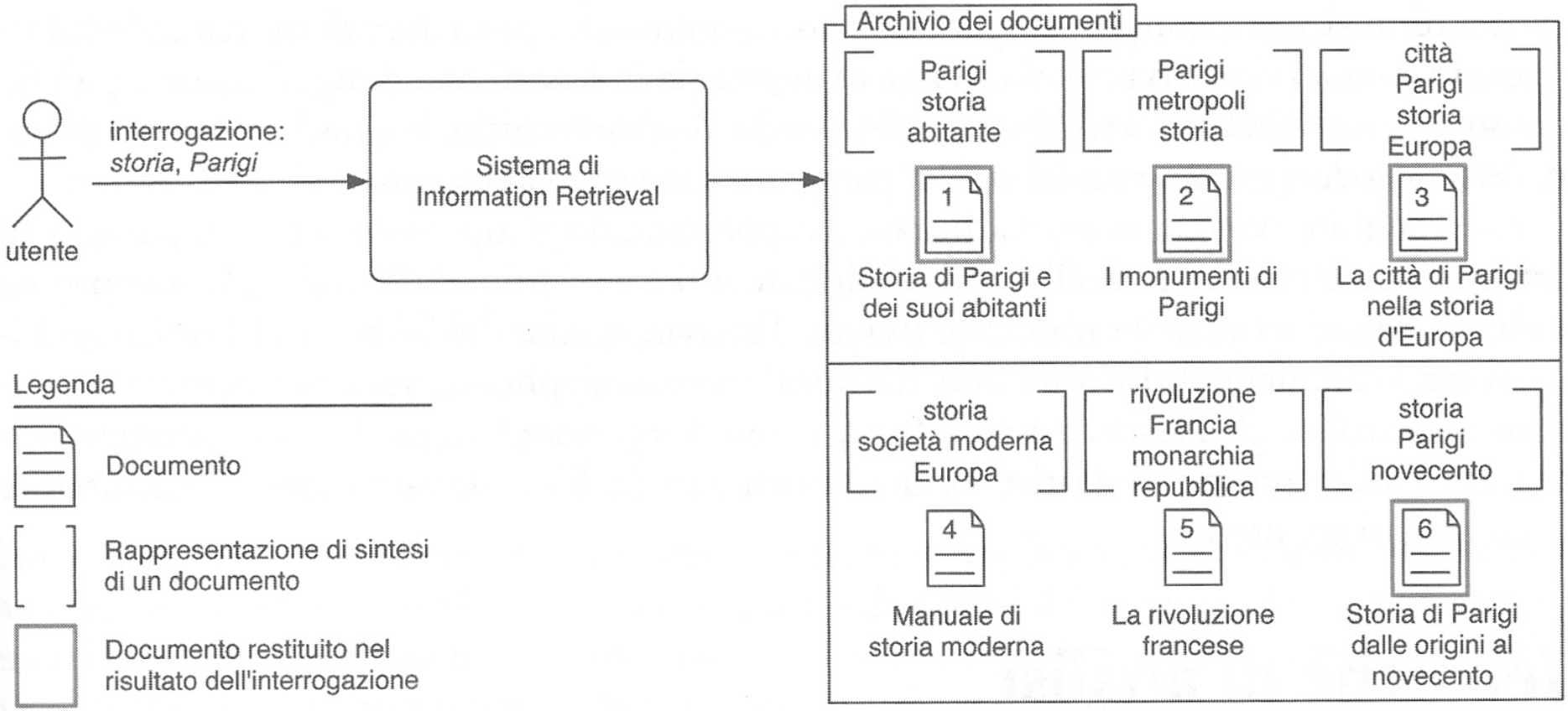
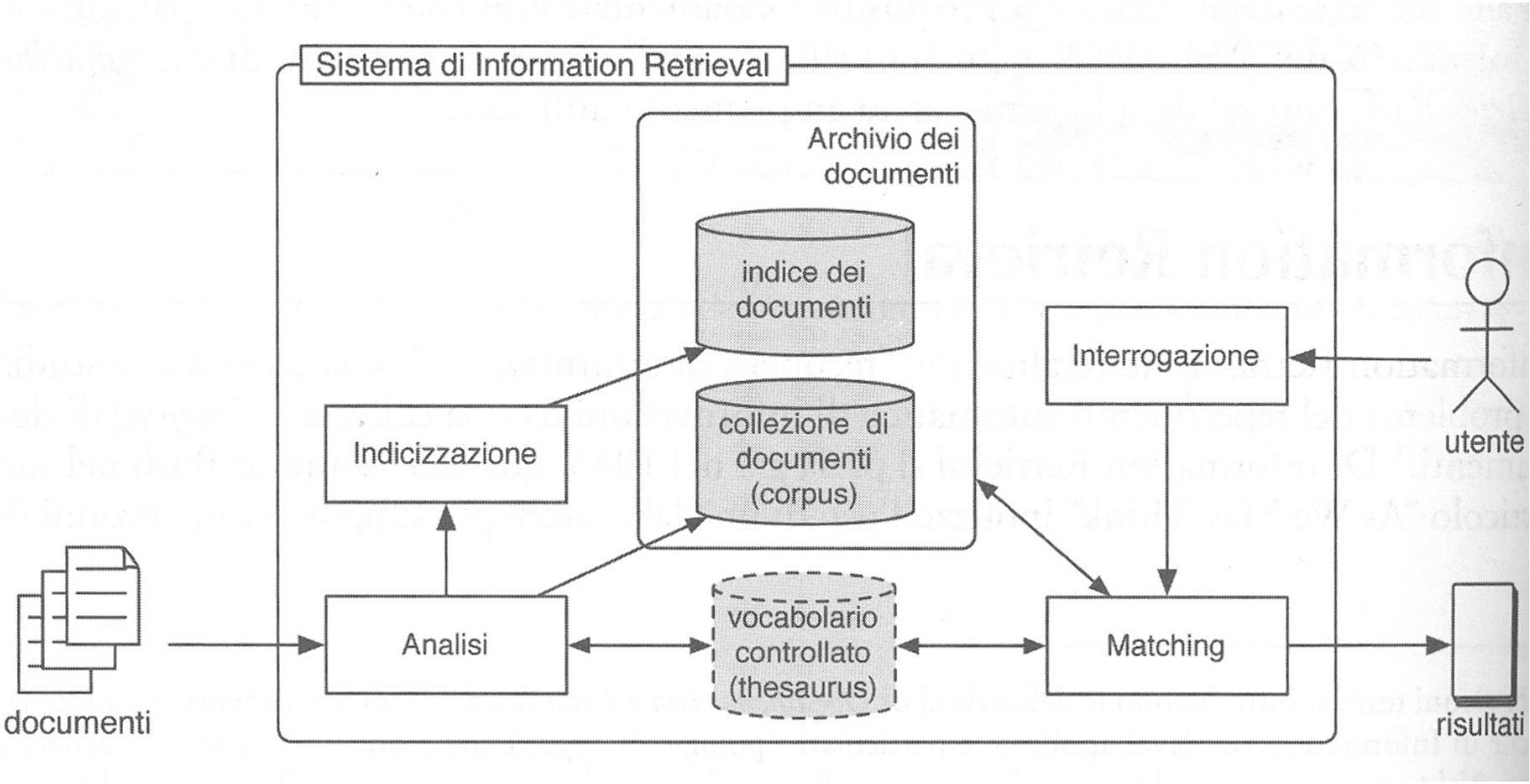


Figura 7.2 Esempio di Information Retrieval.

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.

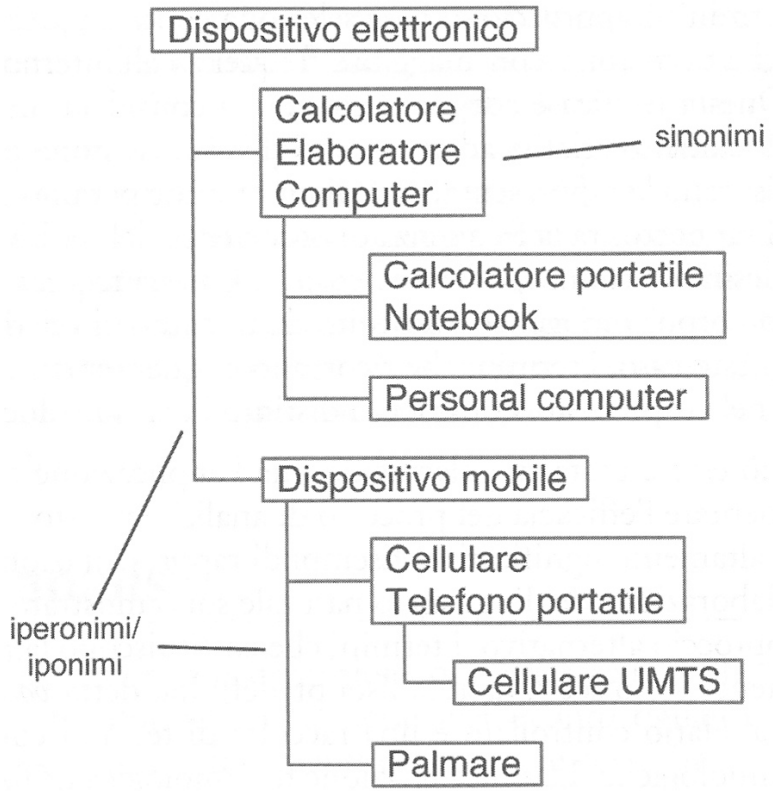
Interrogazione e *matching* ☞"far combaciare"

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*.

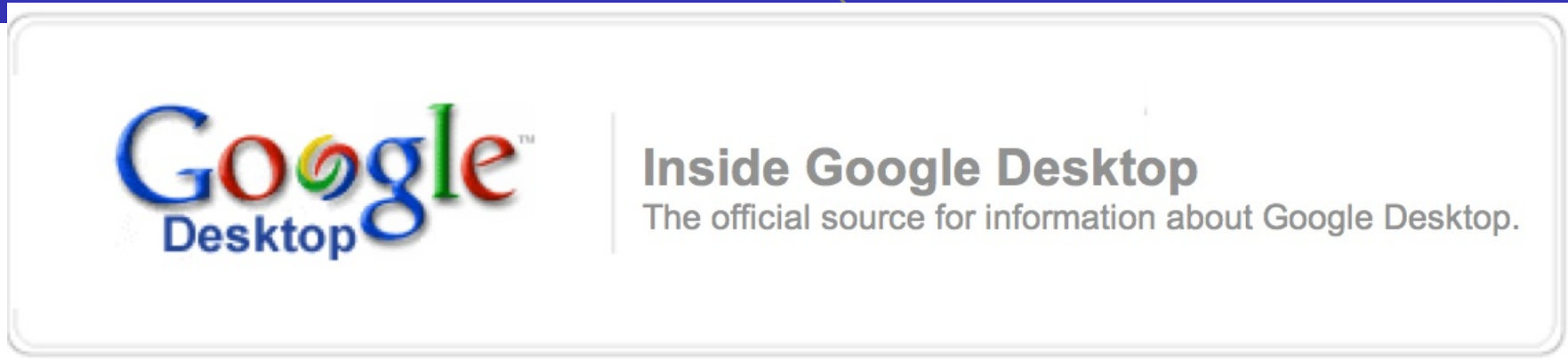
Thesaurus

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)



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,[2] nel senso della morfologia, rispettivam. cosí:

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

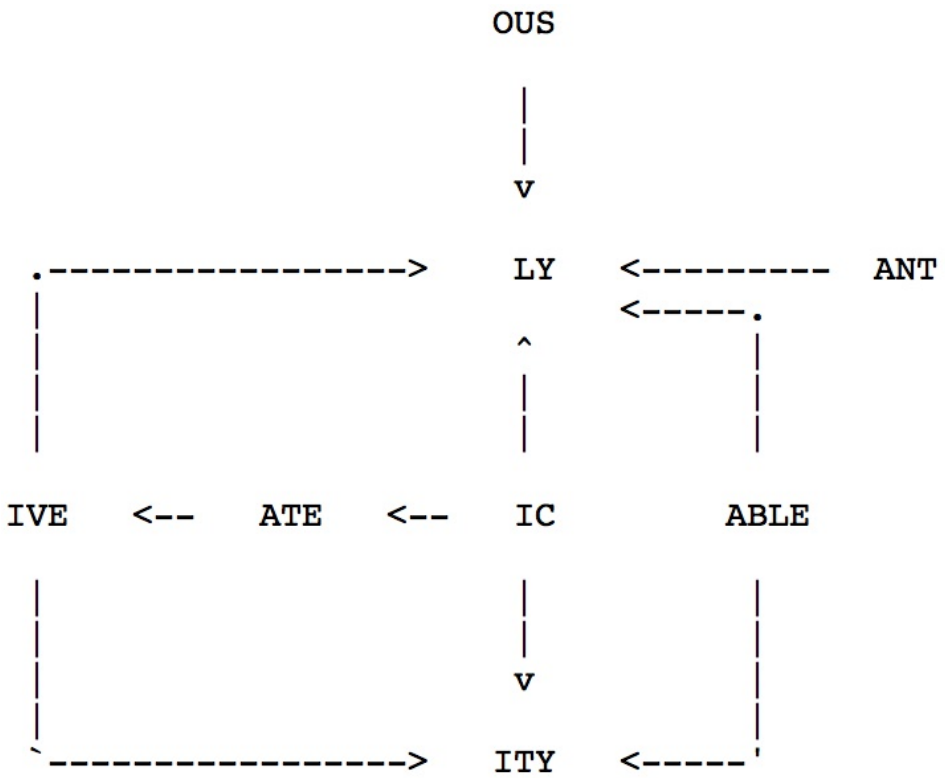
Il processo di *stemming* consiste nel sostituire tutte le forme derivate con il tema corrispondente, per favorire l'estrazione di parole ad elevata rappresentatività.

[2] 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	*ance*	*anza*	*eza*	*anza*
adjective	IC	*ique*	*ico*	*ico*	*ico*
noun	ISM	*isme*	*ismo*	*ismo*	*ismo*
adjective	ABLE	*able*	*able*	*ável*	*abile*
adjective	IBLE	-	*ible*	*ível*	*ibile*
noun	IST	*iste*	*ista*	*ista*	*ista*
adjective	OUS	*eux*	*oso*	*oso*	*oso*
noun	MENT	*ment*	*amiento*	*amento*	*mente*
noun	ATOR	*ateur*	*ador*	*ador*	*attore*
noun	ATRESS	*atrice*	-	-	*atrice*
noun	ATION	*ation*	*ación*	*ação*	*azione*

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.

Strumenti lessicali utili nella formulazione di un thesaurus

ItalWordNet

Parola: motivo Invia Mostra tutte le relazioni ☑

---

motivo, Nome

- [1] - la causa, il principio, l'origine di una cosa o di un'evento; il motivo, la ragione.
  (*cagione* [1], *causa* [1], *motivo* [1], *origine* [3], *perché* [1], *principio* [5], *ragione* [4])
  - has_hyponym **ammennicolo** [1] **anima** [1] **animo** [1] **argomento** [3] **cavillo** [1] **fattore** [3] **favilla** [3] **genesi** [2] **germe** [2] **madre** [5] **movente** [1] **pretesto** [1] **scintilla** [3] **scusa** [4] **seme** [3] **spirito** [1] **uno** [3]
- [2] - idea melodica di un'opera musicale; spunto melodico di un'opera musicale
  (*aria* [3], *melodia* [1], *motivo* [2], *tema* [4])
  - has_hyperonym **componimento_musicale** [1] **composizione** [4] **composizione_musicale** [1] **musica** [2]
  - has_hyponym **cabaletta** [1] **cavatina** [1] **leitmotiv** [2] **motivo** [3]
  - xpos_fuzzynym **armonico** [1] **armonioso** [1] **melodioso** [1] **musicale** [2]
- [3] - motivo facile ed orecchiabile
  (*motivo* [3])
  - has_hyperonym **aria** [3] **melodia** [1] **motivo** [2] **tema** [4]
  - has_hyponym **jingle** [1] **motivetto** [1] **sigla** [2]
- [4] - tema di un'opera letteraria; motivo conduttore
  (*leitmotiv* [1], *motivo* [4], *tema* [3])
  - has_hyperonym **concetto** [1] **concezione** [1] **idea** [1] **pensiero** [2]
  - has_hyponym **head-line** [1]
- [5] - elemento ricorrente nelle arti figurative
  (*motivo* [5])
  - has_hyperonym **figura** [2] **icona** [1] **immagine** [1]

Indicizzazione

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

$\boxed{(\ell_i , R_i)}$ 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, Ray Henry Henry, and Alessandro Amoroso.
FLUENCY –Conoscere e usare l'informatica.
Pearson Italia, Milano-Torino, 7ª edition, 2020.