Matematica e dintorni: storie di contaminazione negli ultimi due secoli

Siena 5-7 aprile 2019 - MATEPRISTEM

La applicazioni informatiche che ci hanno cambiato la vita.

Linda Pagli,
Dipartimento di informatica
Università di Pisa

www.di.unipi.it/~pagli

Applicazioni

- i motori di ricerca: Google
- le reti sociali
- i navigatori
- la crittografia
- i sistemi di raccomandazione: Netflix
- le reti neurali

Come fa a rispondere in un baleno?

Classifica Google 2017

Categoria: come fare a ...

Dal sito Google Trends
http://www.google.com/trends/topcharts

- Le olive in salamoia.
- Il back up.
- La marmellata di albicocche.
- La carbonara.
- Lo screenshot.
- Il pesto
- La crema pasticcera.
- Le bolle di sapone.
- Il passaporto.
- Il cubo di Rubik.

Perché Google è il motore di ricerca più popolare?

Perchè è così veloce a reperire le informazioni nel mare del web?

Come fa a dirci proprio quello che cerchiamo?

Accesso al web prima di Google

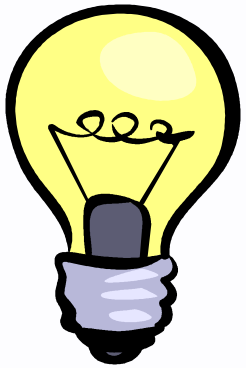
- Sommersi in un mare di informazioni!

Poi arrivò Google

- La ricerca è facile
- La risposta è immediata e quasi sempre soddisfacente.
- L'interfaccia è semplice e pulita senza pubblicità.

Sergey Brin & Larry Page

Parte da un garage e da due studenti di dottorato dell'università di Stanford .



A ogni pagina è associato voto: il page rank

Page rank

- È espresso dal dinamismo della rete, non deciso a priori.
- La risposta del motore è un elenco di pagine ordinate in ordine decrescente di page rank.
- Non giudica sulla reale qualità delle pagine.
- In moltissimi casi funziona bene!

Google è un sistema molto complesso di hardware e software!

Gli ingredienti del successo:

- Web crowling
- Dizionari e indici
- Calcolo parallelo e distribuito
- Page rank

Web Crawling

- Crawler
- Spyder
- Robot

"All Together Now"	1967	*Yellow Submarine*
"All You Need Is Love"	1967	*Magical Mystery Tour*
"And I Love Her"	1964	*UK: A Hard Day's Night* *US: Something New*

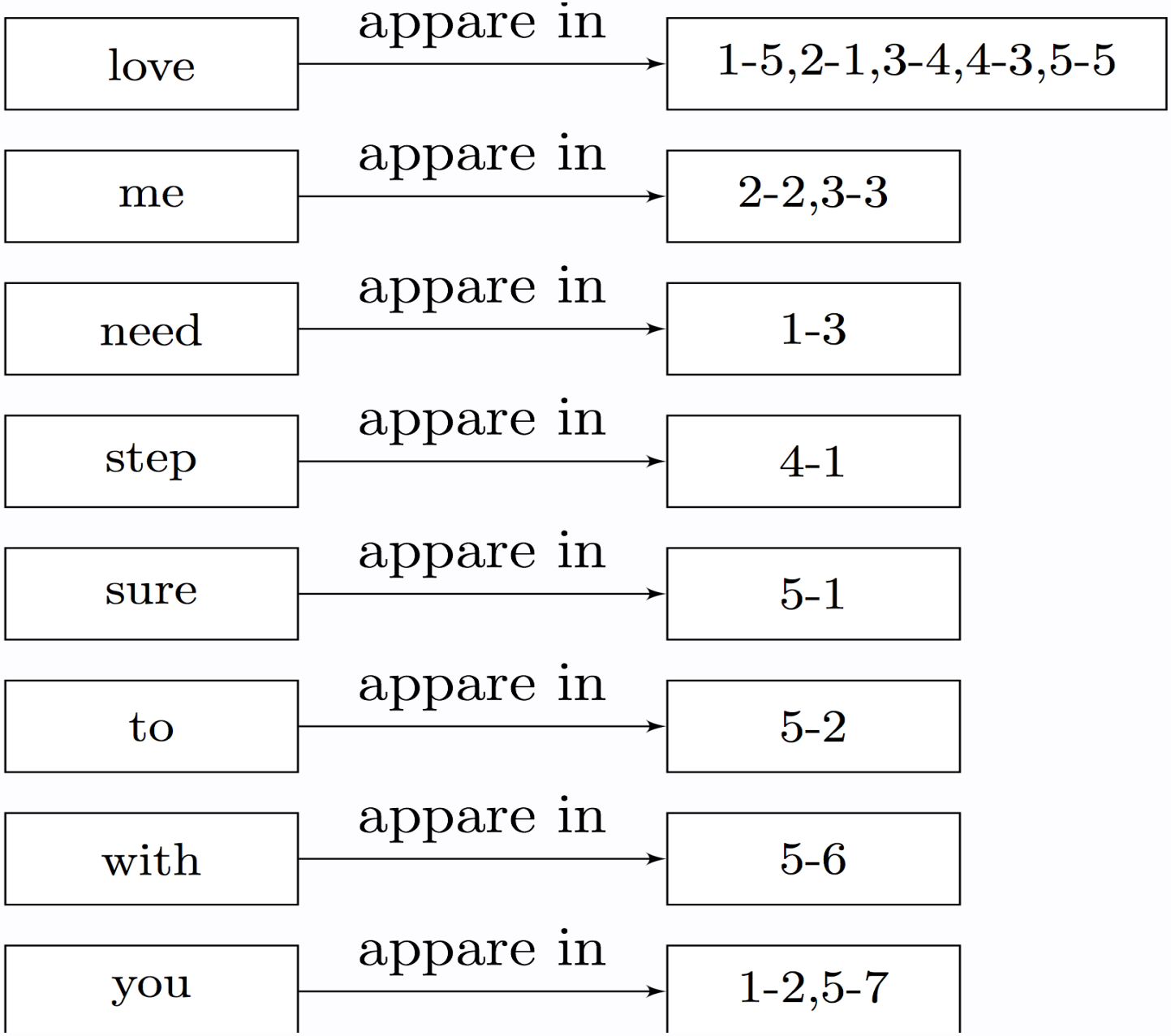
Gli enormi dizionari

Google costruisce un indice analitico di tutte le parole incontrate nello scandaglio continuo delle pagine web.

Un micro-web di 5 pagine

Ogni pagina il titolo di una canzone dei Beatles:

1. All You Need Is Love.
2. Love Me Do.
3. Can't Buy Me Love.
4. Step Inside Love.
5. Sure to Fall (In Love With You).


love
appare in
1-5,2-1,3-4,4-3,5-5
me
appare in
2-2,3-3
need
appare in
1-3
step
appare in
4-1
sure
appare in
5-1
to
appare in
5-2
with
appare in
5-6
you
appare in
1-2,5-7

L'indice

- Le parole sono ordinate in ordine alfabetico per le ricerche veloci.

- Love appare in tutte le pagine. Si indica numero di pagina e posizione:

  L'elenco è 1-5, 2-1, 3-4, 4-3, 5-5.

Ricerca per parole chiave

- 2 parole chiave love e me
- Love: 1-5, 2-1, 3-4, 4-3, 5-5
- Me: 2-2, 3-3

Algoritmo di Merge

Seleziona le pagine comuni: 2 e 3

Ricerca di "Love me"

Pagina 2: Hit!

Pagina 3: prima me poi love, la distanza è maggiore (la massima possibile).

Si cerca solo nell'indice

- La ricerca degli hits si attua nell'indice soltanto. Solo le pagine del risultato dovranno essere reperite nel dizionario.
- Non si devono ricaricare o analizzare le pagine.

- Tecniche note prima di Google

Milioni di richieste al minuto

- Gestite dai suoi famosi data center dislocati geograficamente in tutto il mondo. Composti da cluster.

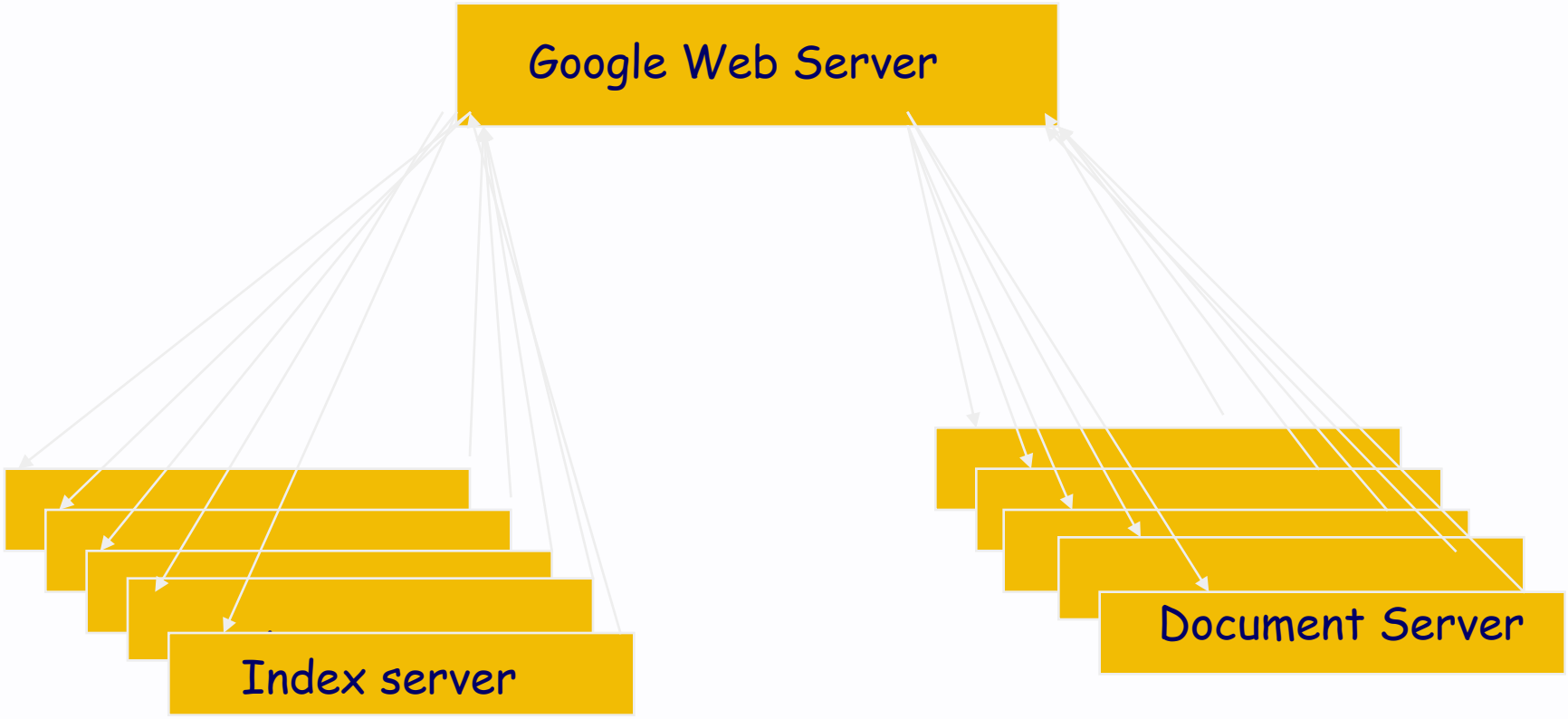
- Un cluster è composto da migliaia di computer (semplici PC) e contiene varie repliche di tutto il web.

Una singola richiesta (query)

- una query a Google in media:
  - legge centinaia di Megabytes di dati
  - consuma decine di miliardi di cicli di CPU

- Google gestisce milioni di queries/sec
- La nostra query viene smistata al data center più vicino o, se è molto occupato, a quello più sgombro

- Nel cluster:



Per ogni richiesta un GWS deve:

- Avviare la ricerca delle parole chiave negli indici e restituire i riferimenti alle pagine web che contengono le parole chiave (hit list)
- Intersecare le hit list per ciascuna parola chiave per determinare i documenti rilevanti e reperirli.
- Calcolare il voto (rank) di ciascuna pagina per ordinare il risultato.

Ricerca: 2 livelli di parallelismo

La parola da cercare è replicata in tante copie.

1. L'indice è diviso in pezzi, ciascuno gestito da un insieme di macchine (index servers) per la ricerca parallela della parola sul pezzo.
2. La parola viene mandata in parallelo a tutti i pezzi.

- Risposta:

Recupero dei documenti della risposta nell'archivio che contiene la copia completa del web a disposizione del cluster.

Come nella fase della ricerca i document server effettuano la ricerca contemporanea dei
dei documenti e in parallelo sui pezzi.

In questa fase si assegnano ai documenti: titolo, Riassunto e un estratto che mostra la parole chiave cercata nel contesto del documento.

Page rank

- Come fare le olive in salamoia. 2 pagine web danno la ricetta

"giallozafferano"

"sale&pepe"

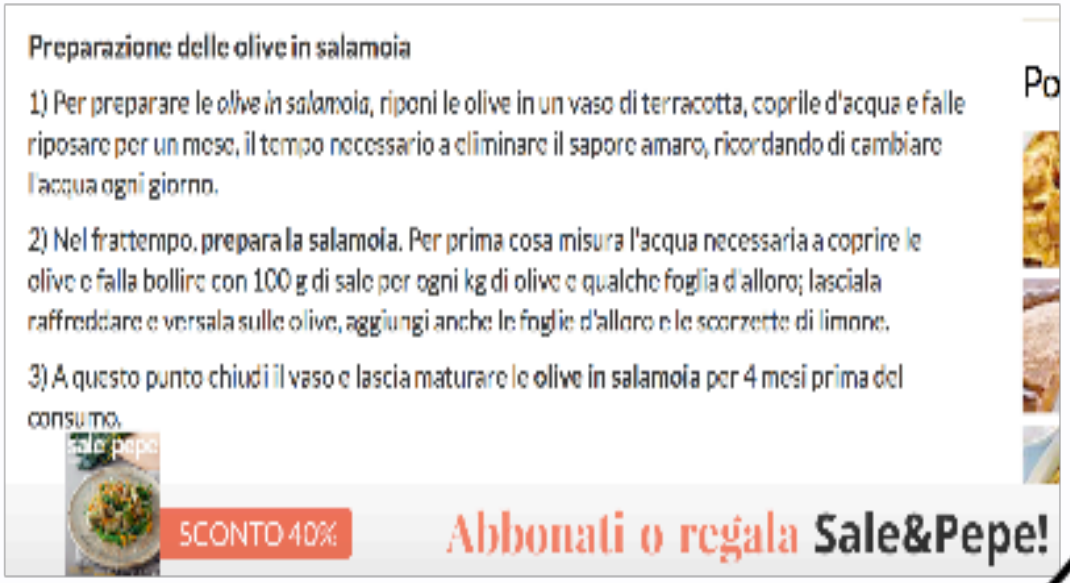
Strategia immediata per dare il voto:
Contare il numero di riferimenti da altre pagine alla pagina in questione.

Page rank

Se contiamo il numero di riferimenti le due pagine coincidono. Ma una cosa è la menzione di un tizio qualsiasi altra cosa la menzione di un personaggio famoso.

Deve contare la popolarità (rank) della pagina che contiene la citazione e quante citazioni fa.


Il diario delle mie ricette
Olive verdi in salamoia


Preparazione delle olive in salamoia
1) Per preparare le olive in salamoia, riponi le olive in un vaso di terracotta, coprile d'acqua e falle riposare per un mese, il tempo necessario a eliminare il sapore amaro, ricordando di cambiare l'acqua ogni giorno.
2) Nel frattempo, prepara la salamoia. Per prima cosa misura l'acqua necessaria a coprire le olive e falla bollire con 100 g di sale per ogni kg di olive e qualche foglia d'alloro; lasciala raffreddare e versala sulle olive, aggiungi anche le foglie d'alloro e le scorzette di limone.
3) A questo punto chiudi il vaso e lascia maturare le olive in salamoia per 4 mesi prima del consumo.
SCONTO 40%
Abbonati o regala Sale&Pepe!


facebook
Chef Carlo Cracco

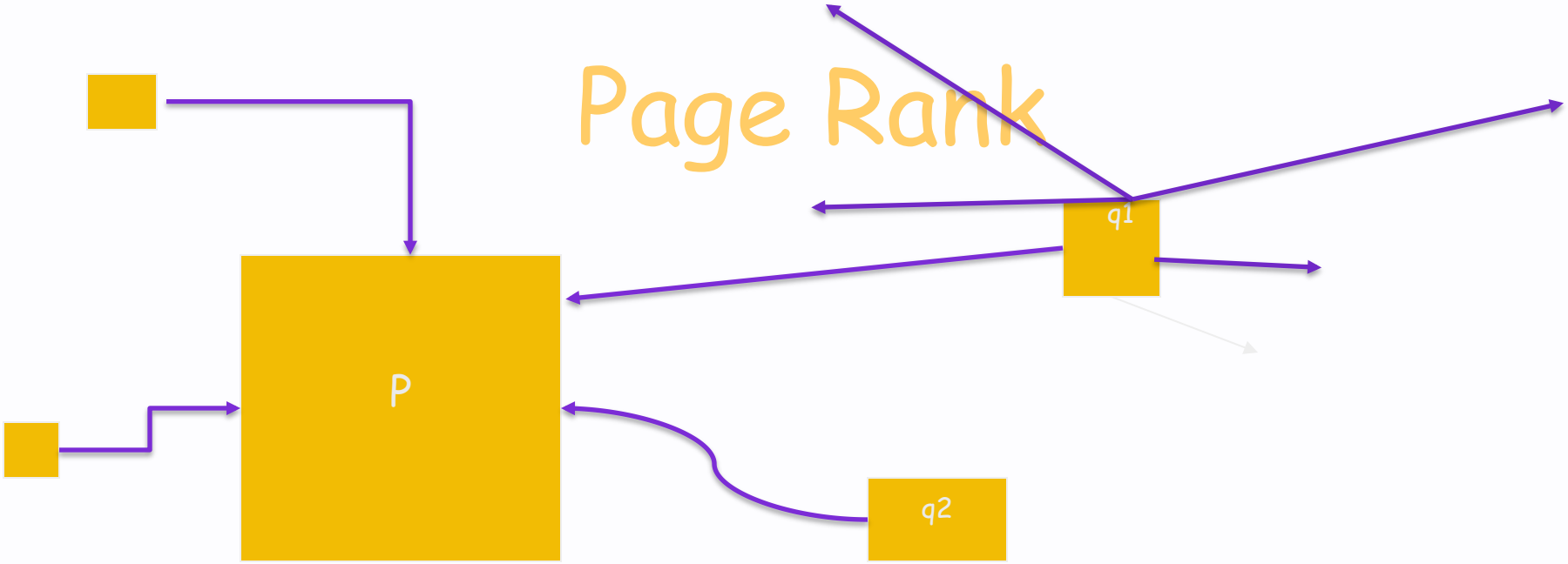
Mario

Calcolo del page rank

R(Sale&pepe)= rank(Mario)/#link(Mario)=1/2

R(giallozafferano)= rank(chef)/#link(chef)=100/100 =1

Page Rank



q cita p

$$\text{Page Rank di P} = \sum \frac{\text{Page Rank di } q}{\text{numero di pagine citate da q}}$$

$$= \text{rank}(q_1)/5 + \text{rank}(q_2)/1 + \cdots$$


PageRank

Importante è la popolarità sul web non l'autorevolezza

- Nel nostro esempio il personaggio famoso è anche uno chef, quindi autorevole, ma poteva essere anche un calciatore famoso e il risultato sarebbe stato uguale.

Il Page Rank non è l'unico parametro considerato

- Altri tenuti rigorosamente segreti.

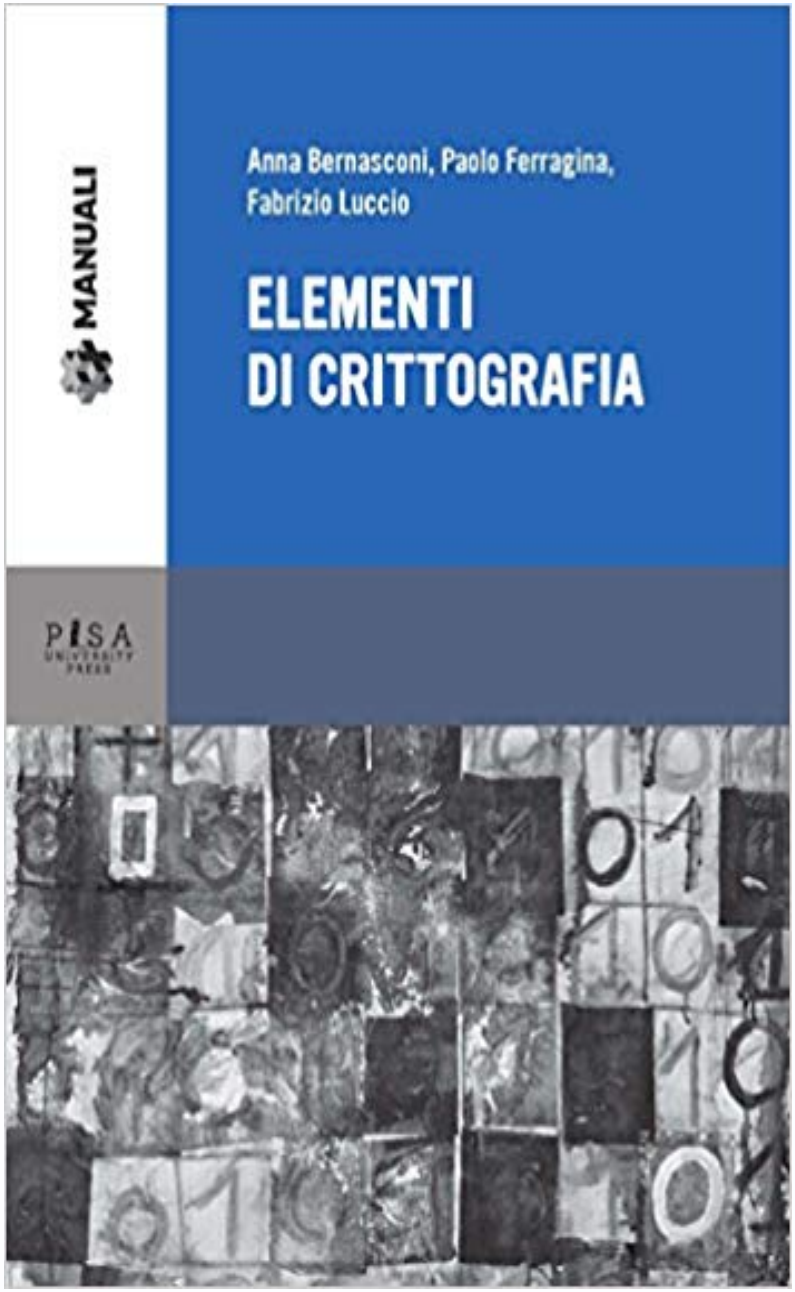
- Probabilmente:
  - Le home page hanno rank più alto
  - Il fatto che una pagina sia fresca
  - ….molti altri

Pierluigi Crescenzi
Linda Pagli

PROBLEMI, ALGORITMI E CODING

Le magie dell'informatica

ZANICHELLI


MANUALI
Anna Bernasconi, Paolo Ferragina,
Fabrizio Luccio
ELEMENTI
DI CRITTOGRAFIA
PISA
UNIVERSITY
PRESS