INFORMATICA PER LA COMUNICAZIONE

MOTORI DI RICERCA

Motori di ricerca

Si stima vi siano almeno **diversi miliardi** di pagine web, oltre **un miliardo** di siti web

Sistemi per la ricerca e la catalogazione delle pagine web

- **motori di ricerca** (Google, Yahoo!, Bing,...)
- **cataloghi sistematici** (Yahoo! Directory, Open Directory Project)

Motori di ricerca

- **Motore di ricerca:** strumento di **IR** (***Information Retrieval)***
- Ricerca per **parole chiave**
- **Risultato di una ricerca: lista ordinata** di pagine che trattano argomenti descritti dalle parole chiave
- Presenza collegamenti sponsorizzati

Motori di ricerca

- Numero interrogazioni: migliaia al secondo
- **Tempi di risposta istantanei**

Informazione e Web

Caratteristiche del Web

- Grande collezione di documenti
- Informazione **non strutturata**, priva di **organizzazione**
- Tecniche delle **basi di dati** → **non applicabili** direttamente al Web
- Tecniche alternative di **indicizzazione** e interrogazione

Information Retrieval

- ***Information Retrieval (IR)***: disciplina che studia come rispondere alle **esigenze informative**
  - Cataloghi biblioteca
  - Motori di ricerca
  - …
- Processo di selezione dell'informazione **rilevante** all'interno di un **corpus** di documenti

Information Retrieval

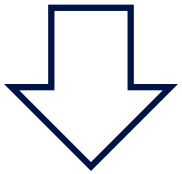
Sistema IR

- **Interrogazione** (*query*) dell'utente: stringa di testo
  - Es. **Parole chiave**
- **Risposta** del sistema
  - **Selezione** all'interno del corpus della parte **rilevante**
  - Eventuale **ordinamento** risultato

Information Retrieval

Obiettivi sistemi IR

- **Efficacia**: accuratezza della risposta
- **Efficienza**: velocità nel fornire risposte



Obiettivi contrastanti

Information Retrieval

Sistemi IR analizzano **preventivamente** documenti

- Estrarre **informazione significativa**
- Memorizzazione **efficiente**

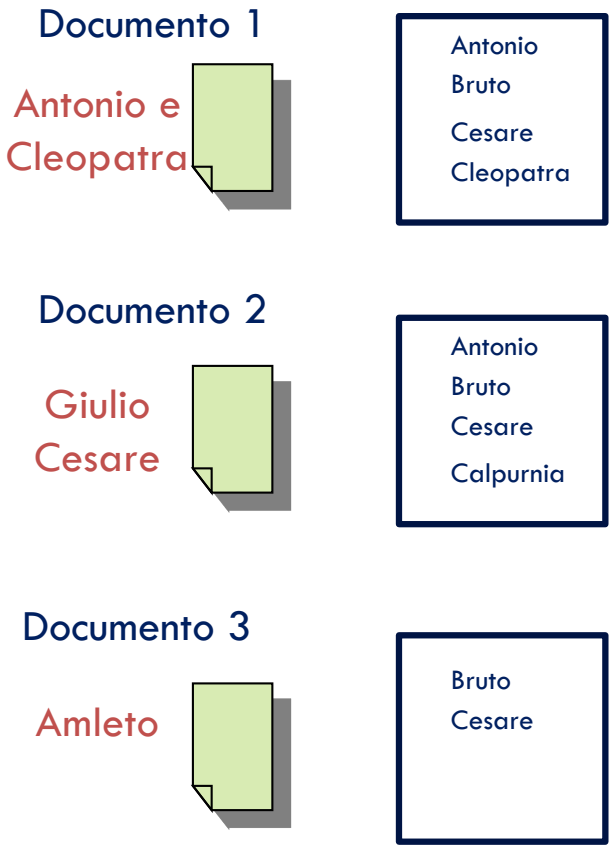
Obiettivi:

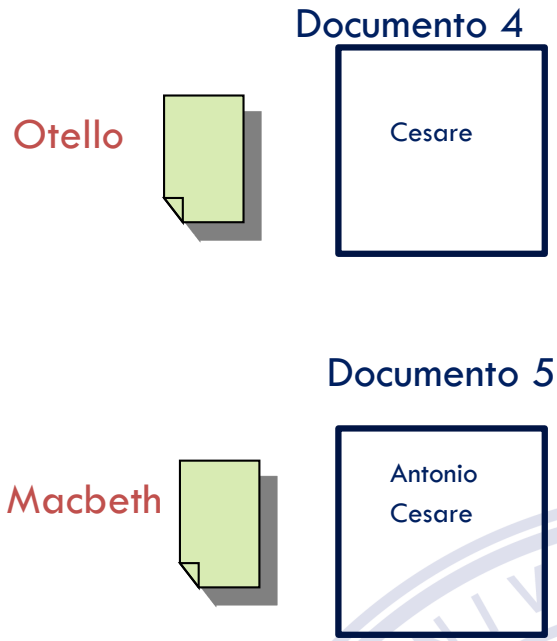
- Rilevanza risultati
- Limitare i confronti

Tecniche di indicizzazione

- **Indice**: struttura per reperimento **efficiente** di **documenti correlati a interrogazione**
- Riduzione tempi di ricerca
- Informazioni memorizzate nell'indice
  - **Documenti** in cui compare T
  - **Frequenza** T
  - **Parte** in cui compare T

Information Retrieval





Tecniche di indicizzazione

- **Struttura di un indice**
  - Termine T
  - Riferimenti ai documenti collegati a T

Termine	Riferimenti
Antonio	Documento 1 (frequenza 125); Documento 2 (frequenza 43), Documento 5 (frequenza 3)
Bruto	Documento 1 (frequenza 15); Documento 2 (frequenza 83), Documento 3 (frequenza 2)
Cesare	Documento 1 (frequenza 115); Documento 2 (frequenza 298), Documento 3 (frequenza 2), Documento 4 (frequenza 2), Documento 5 (frequenza 3)
Calpurnia	Documento 2 (frequenza 21)
Cleopatra	Documento 1 (frequenza 124)

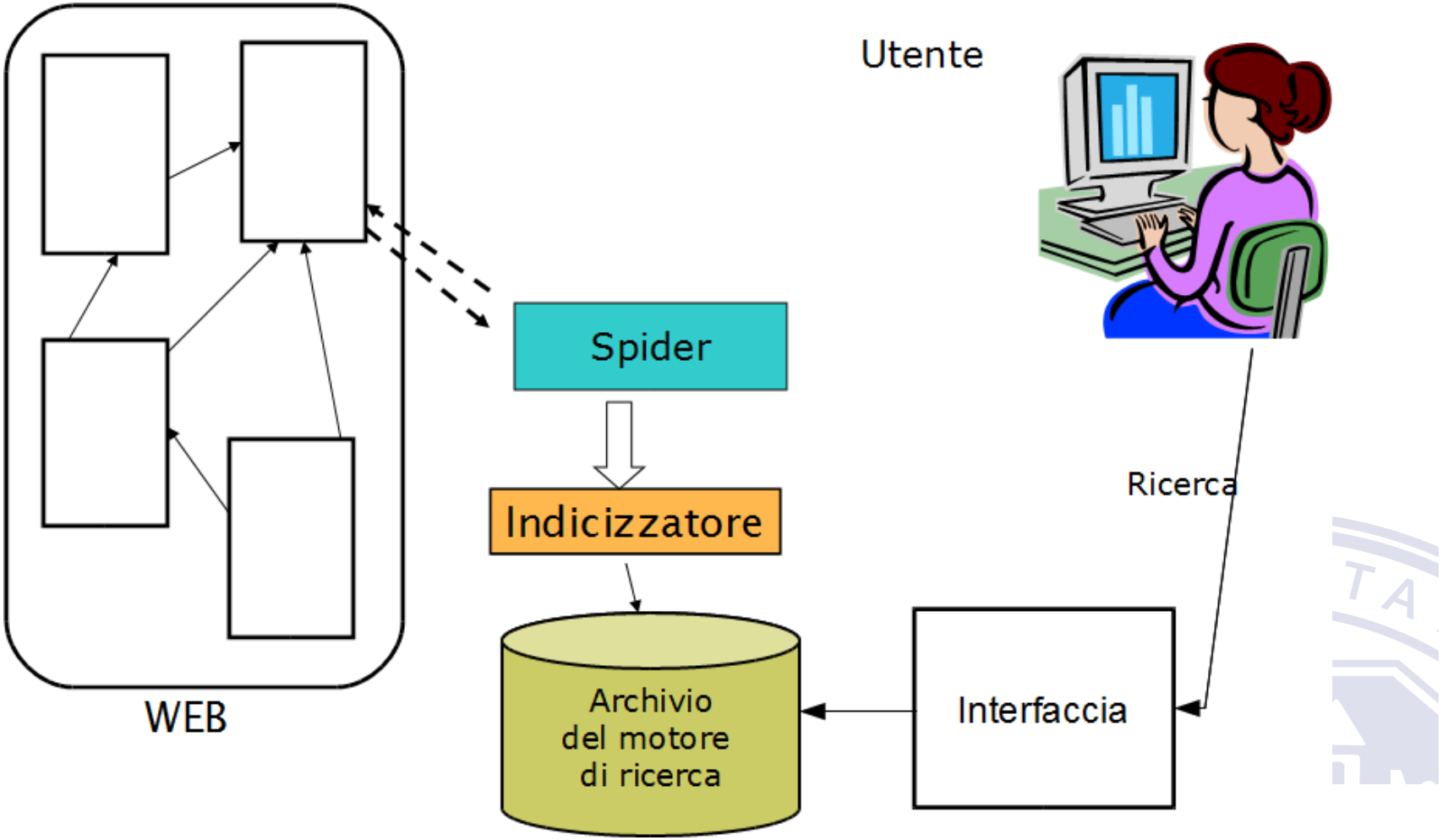
Interrogazione

- **Risultato interrogazione**:
  - Lista documenti **correlati alla ricerca**
  - **Valore di rilevanza** → similarità tra parole chiavi e termini documento
- Documenti con scarsa rilevanza → scartati

Motori di ricerca

- Come funziona un motore di ricerca?
- **Attività** di un motore di ricerca:
  - **Raccolta**
  - **Analisi e indicizzazione**
  - **Interrogazione**

Motori di ricerca



Motori di ricerca

Raccolta: *esplorazione* del Web

- Utilizzo link
- Generazione indirizzi
- Programmi per la raccolta: ***crawler, spider, robot***
- Informazioni raccolte memorizzate in base di dati
  → **archivio Web**

Motori di ricerca

Analisi e indicizzazione

- Elaborazione delle informazioni raccolte dagli spider ⟶ **contenuto informativo**
- Utilizzo tecniche IR
  - Analisi del documento
  - Costruzione indice

Parole chiave	**Pagine web**
Notizie	www.corriere.it www.repubblica.it …
Università	www.unimi.it www.unibg.it …
…	…

Motori di ricerca

Interrogazione: risposta a richieste utenti

- **Selezione** dei documenti rilevanti
- Accesso solo all'**indice** e all'**archivio** del motore di ricerca
- Risultati della ricerca → **ordine con cui sono mostrati**
- Pagine ordinate in base a rilevanza (*ranking*)

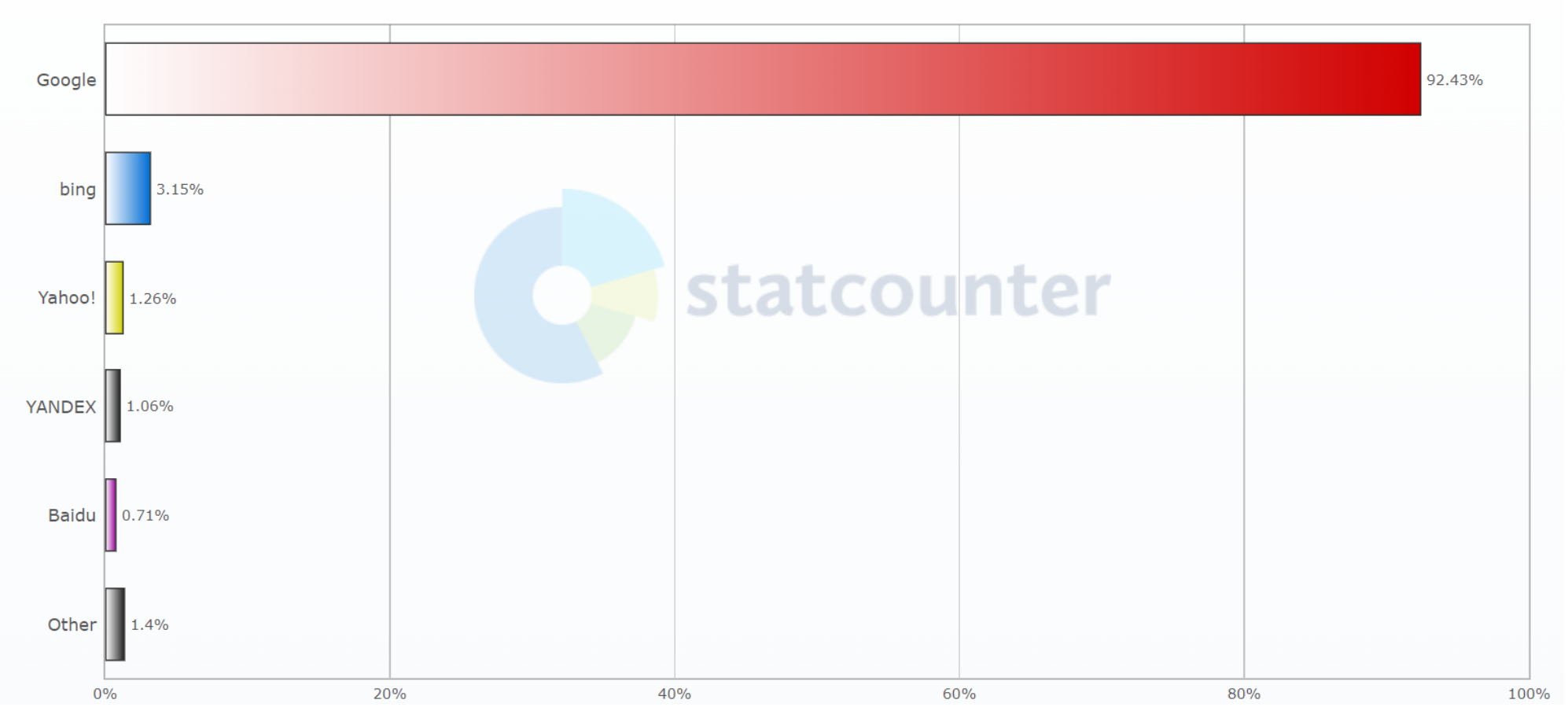
Motori di ricerca

Motori di ricerca **generalisti**

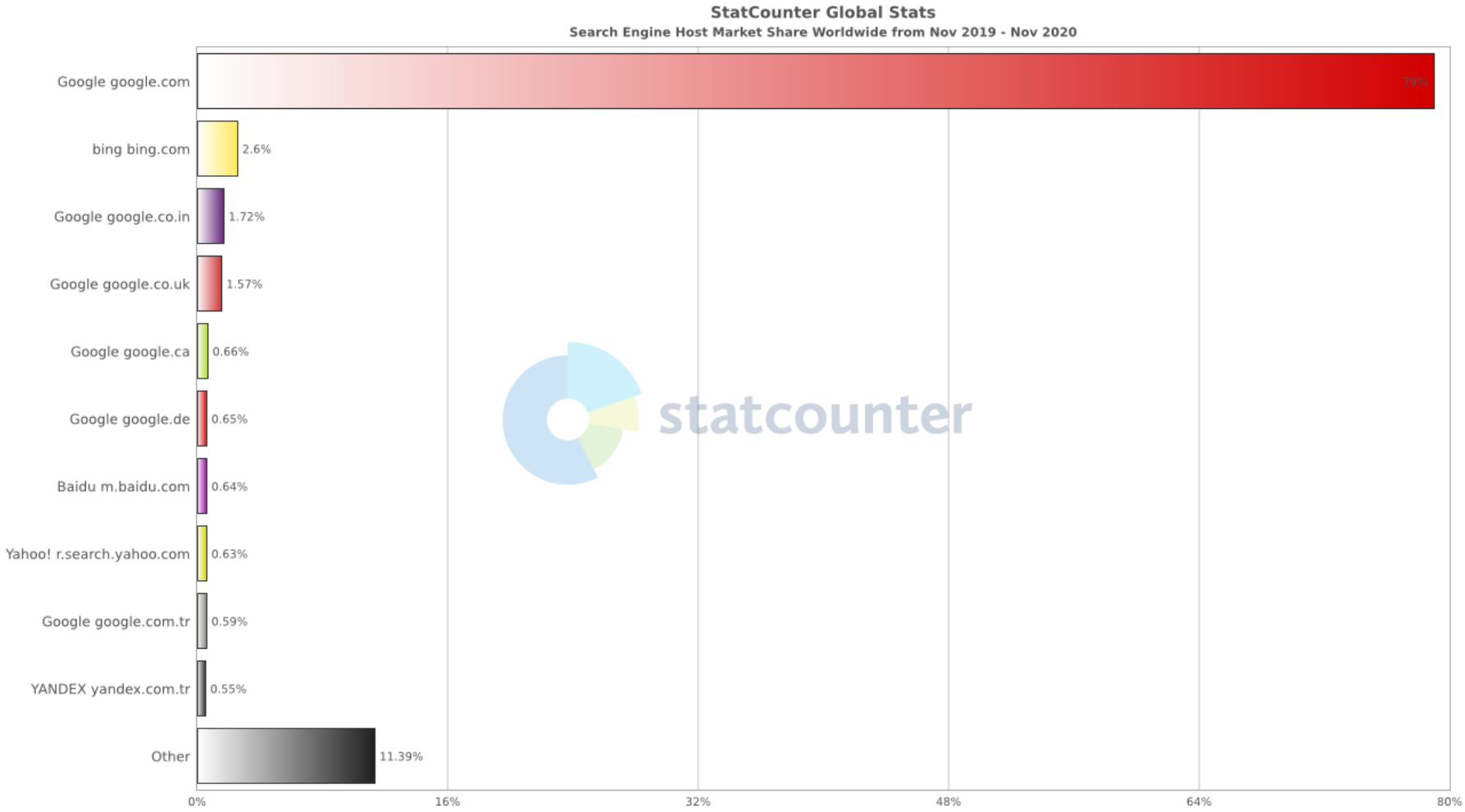
- **Google** (1998)
- **Yahoo! Search** (2004)
- **DuckDuckGo** (2008)
- **Bing** (2009)
- …

Difficoltà nel confronto tra i diversi motori di ricerca

Motori di ricerca



Motori di ricerca



Altri motori di ricerca

- **Metamotori**: sintesi di una ricerca su più motori
- **Plurimotori**: ricerche in parallelo
- **Motori specialistici**: motori di ricerca in ambiti specifici
  - Pubblicazioni scientifiche: Google Scholar
  - Notizie: Google News
  - Libri: Google Books
  - Immagini, video

Motori di ricerca

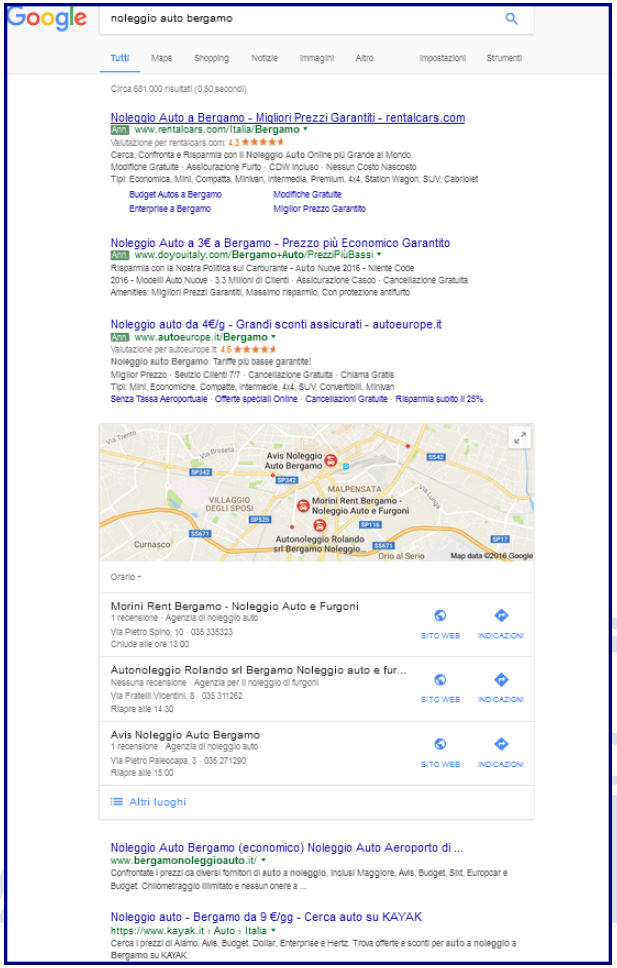
- **Interrogazione**: risposta a richieste utenti
- Accesso solo all'indice e all'archivio Web
- Risultati della ricerca → **ordine con cui sono mostrati**
- Pagine ordinate in base a rilevanza (*ranking*)
- Come funzionano **nella realtà?**
- Come **costruiscono il ranking?**

Ordinamento

Ricerca e ordinamento dei risultati

SERP

- **SERP** (*Search Engine Results Page*): risultati di un motore di ricerca
  - Risultati **organici**
  - Risultati **a pagamento**
  - Ricerca sulle mappe
  - …



Posizionamento

- Un aspetto fondamentale per la **visibilità** di un sito
  → **posizionamento** nella SERP
- Gli utenti usano spesso i motori di ricerca
  - Nel 2017 circa 3,5 miliardi di ricerche al giorno

Importanza del posizionamento

- Da una ricerca di AOL 2006:
  - I **primi 10** risultati: **89.71% delle selezioni**
  - Risultati da 11 a 20: 4.37%
  - Risultati da 21 a 30: 2.42%
- In base alla posizione
  - **42%** delle selezioni: **prima posizione**
  - **12%** delle selezioni: **seconda posizione**
  - **9%** delle selezioni: **terza posizione**
  - **4%** delle selezioni: **quarta posizione**

Visibilità dei risultati

Rank	Visibilità
1	100%
2	100%
3	100%
4	85%
5	60%
6	50%
7	50%
8	30%
9	30%
10	20%

Motori di ricerca

I motori di ricerca costruiscono i risultati da presentare in **due fasi**:

- **Individuazione** delle pagine rilevanti (IR)
- **Ordinamento** (*Ranking*) delle pagine rilevanti

Motori di ricerca

- Motori di ricerca cercano le pagine che contengono le parole chiave inserite
- Come funzionano **nella realtà?**
- Esempio:
  - Esempio1.html
  - Esempio2.html
  - Esempio3.html
  - Esempio4.html
- Ordinamento rispetto alla ricerca: Università Bergamo?

Motori di ricerca

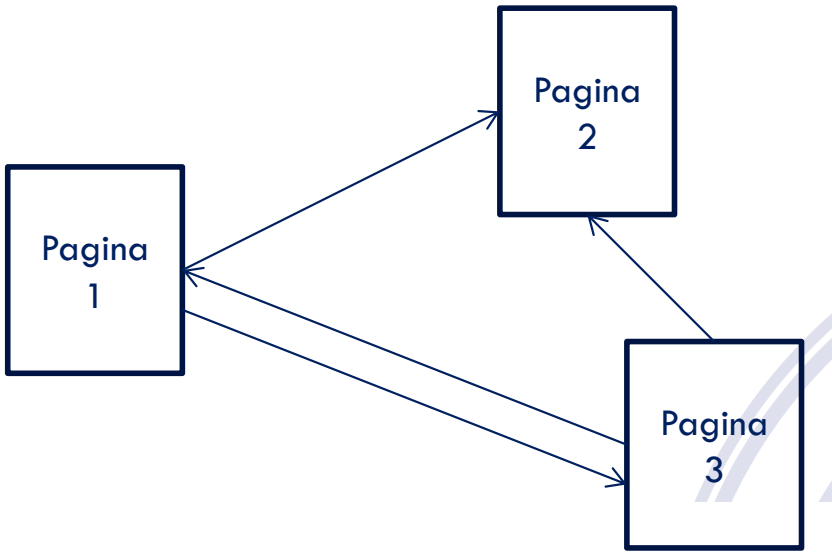
- Alla fine degli anni ’90 analisi della **struttura del Web** per individuare le pagine **più rilevanti**
- Da cosa è costituita questa struttura?
  - Pagine web
  - **Link ipertestuali**
- Link ipertestuali → **legame** tra le pagine web

Grafo del Web

Motori di ricerca

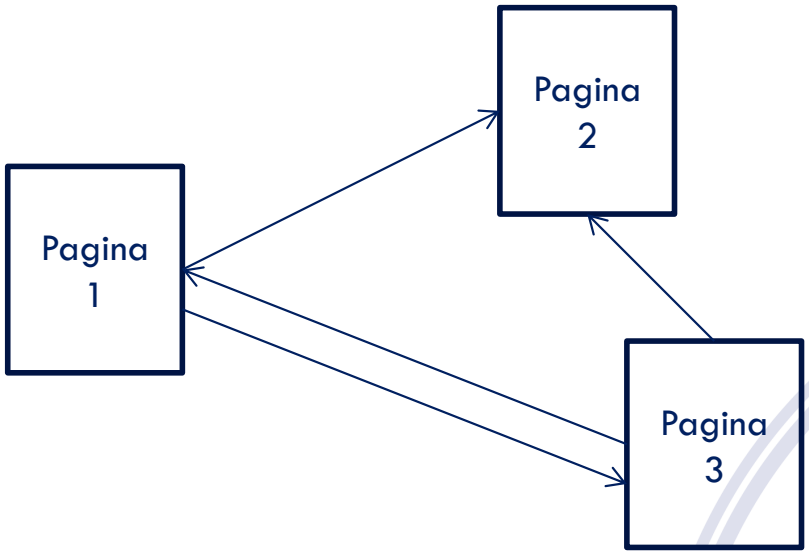
- Alla fine degli anni ’90 analisi della **struttura del Web** per individuare le pagine **più rilevanti**

Grafo del Web



Grafo del Web

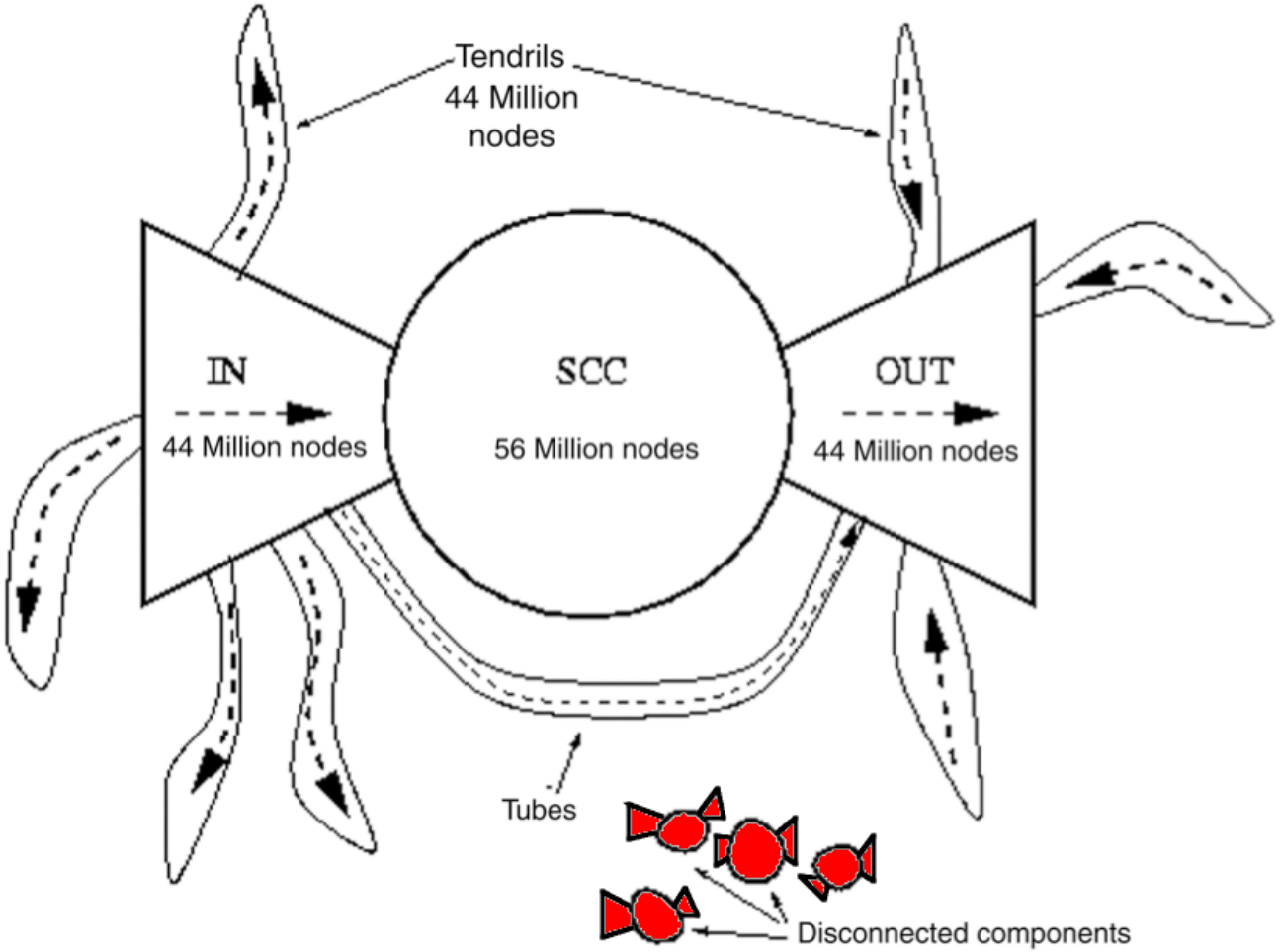
- Grafo del Web
  - **Nodi → pagine web**
  - **Archi (orientati) → link ipertestuali**



Grafo del Web

- Caratteristiche del grafo del Web
  - Rete **orientata**
    - Differenza tra grado uscente e grado entrante
  - **Grado di separazione**: 19 link (*Barabasi*)
  - Presenza di
    - Hub
    - Componenti isolate

Grafo del Web



Grafo del Web

Grafo del Web è utilizzato per

- **Esplorazione** del Web e **raccolta** di dati
- **Ordinamento** dei risultati dei motori di ricerca (*PageRank)*
- Determinare pagine che si occupano di **argomenti simili**

Indicizzazione

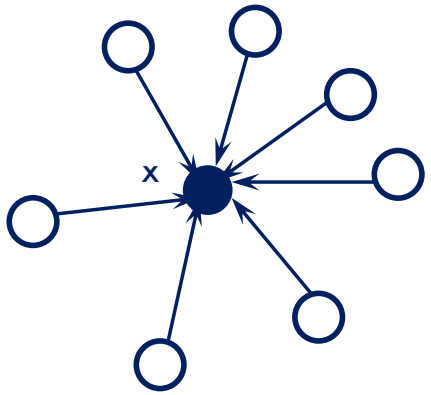
- La struttura del Web condiziona l'**esplorazione** e la conseguente **indicizzazione**
- Nel 1997
  - Hotbot copriva circa il 33% delle pagine
  - Altavista il 28%
  - Lycos 2%

Indicizzazione

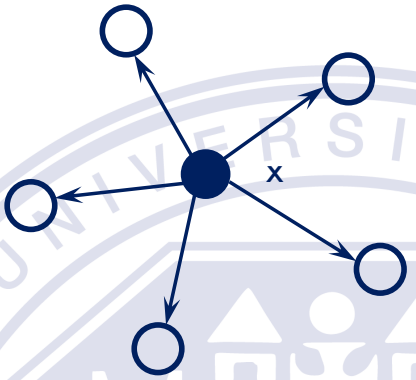
- La probabilità che una pagina sia indicizzata dipende dal **numero di archi entranti**
  - Un link: 10% di probabilità
  - Tra 21 e 100 link: 90% di probabilità

Ranking

- Due caratteristiche di una nodo/pagina web *x :*
- link entranti: **stella entrante**



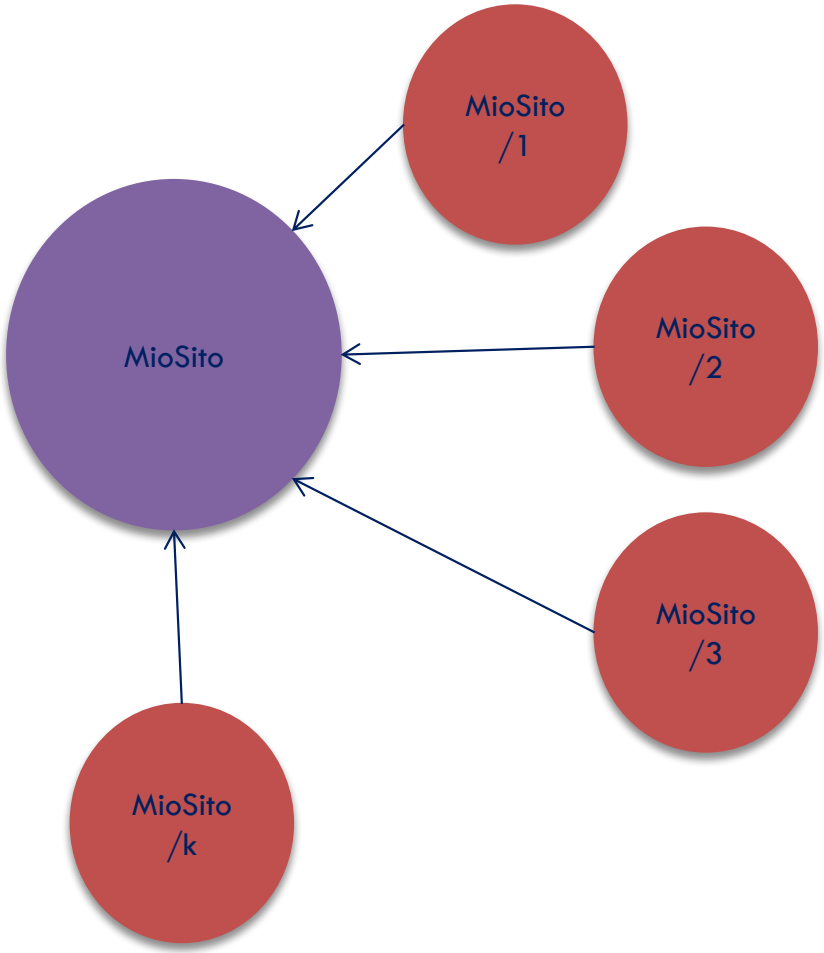
- link uscenti: **stella uscente**



Ranking

- Parametro con cui valutare le caratteristiche di una pagina:
  - **Numero di link** nella stella entrante
  - **Numero di link** nella stella uscente
- **Siti autorità:** siti autorevoli su un certo argomento
  → **molti link entranti**

Ranking



Motori di ricerca

- Ulteriori analisi del grafo del Web per definire migliori criteri di **ordinamento**
- Metodi:
  - *HITS*
  - *PageRank*

HITS

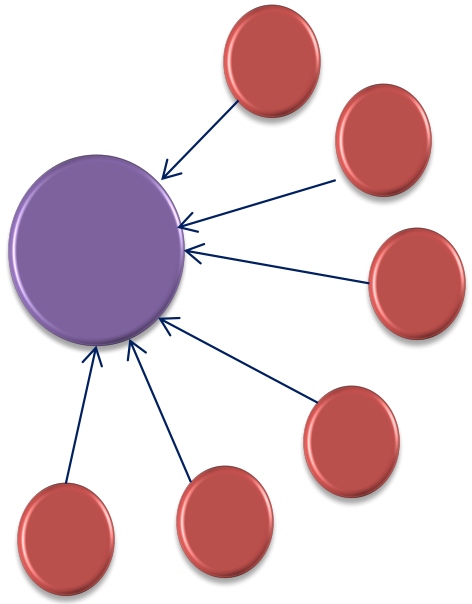
***Hyperlink-Induced Topic Search* (HITS):**

- Analisi di Kleinberg (1999)
- Nodi del grafo del Web possono essere classificati come
  - **Autorità**
  - **Aggregatori**

HITS

***Hyperlink-Induced Topic Search* (HITS):**

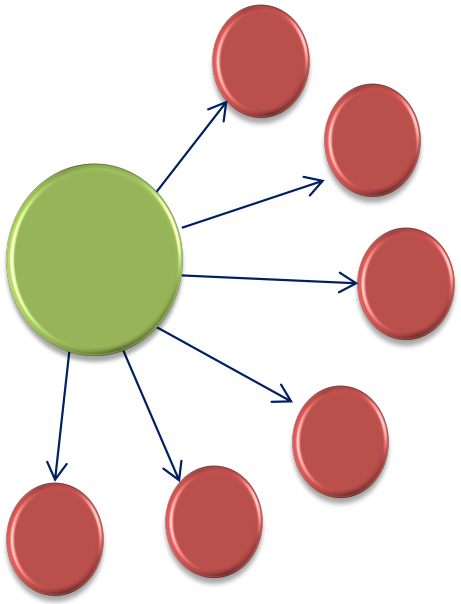
- **Autorità**: nodo con stella entrante molto ampia
  - **Autorevolezza su un dato argomento**



HITS

***Hyperlink-Induced Topic Search* (HITS):**

- **Aggregatore**: nodo con stella uscente molto ampia



HITS

Pagine autorevoli e aggregatori **si influenzano** vicendevolmente

- Aggregatori **rilevanti** → pagine **autorevoli**
- Pagine **autorevoli** → Aggregatori **rilevanti**

PageRank

Successo di **Google** → metodo/algoritmo di **ranking** (PageRank)

*"**The heart of our software is PageRank™**, a system for ranking web pages developed by our founders Larry Page and Sergey Brin at Stanford University. And while we have dozens of engineers working to improve every aspect of Google on a daily basis, **PageRank continues to play a central role in many of our web search tools.**"*

(Google technology)

PageRank

- Ranking di Google basato su ***PageRank***
- **Rilevanza di una pagina web**: funzione della **rilevanza delle pagine web collegate** (rilevanza pagine delle stelle entranti)

PageRank

- Calcolo *PageRank*: **procedura *ricorsiva*** per il calcolo della rilevanze delle pagina
  - Rilevanza di una pagina web **dipende dalla rilevanza** delle pagine web **della stella entrante**
  - Rilevanza della delle pagine web nella **stella uscente influenzata dalla rilevanza mia pagina web**

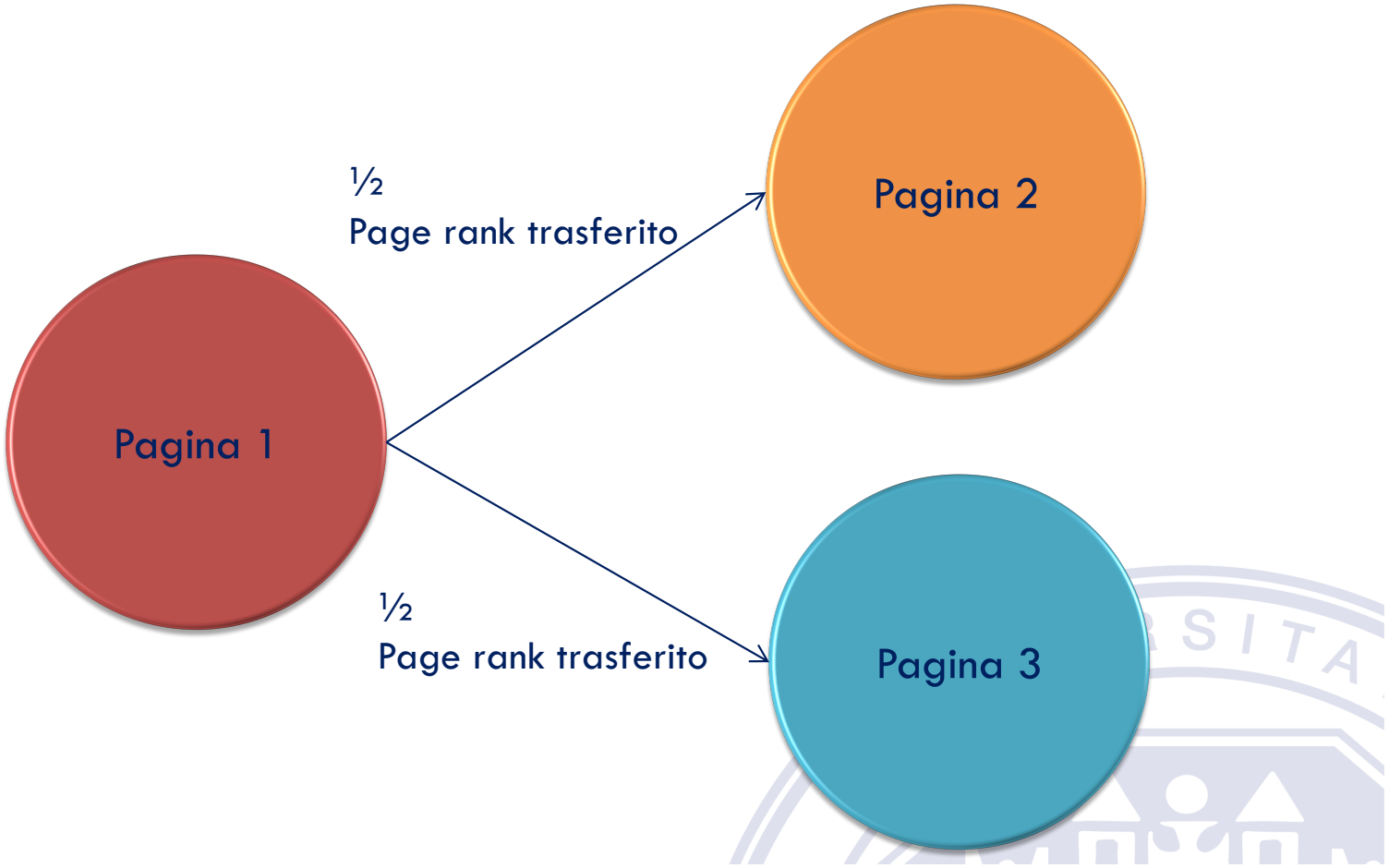
PageRank

Intuitivamente:

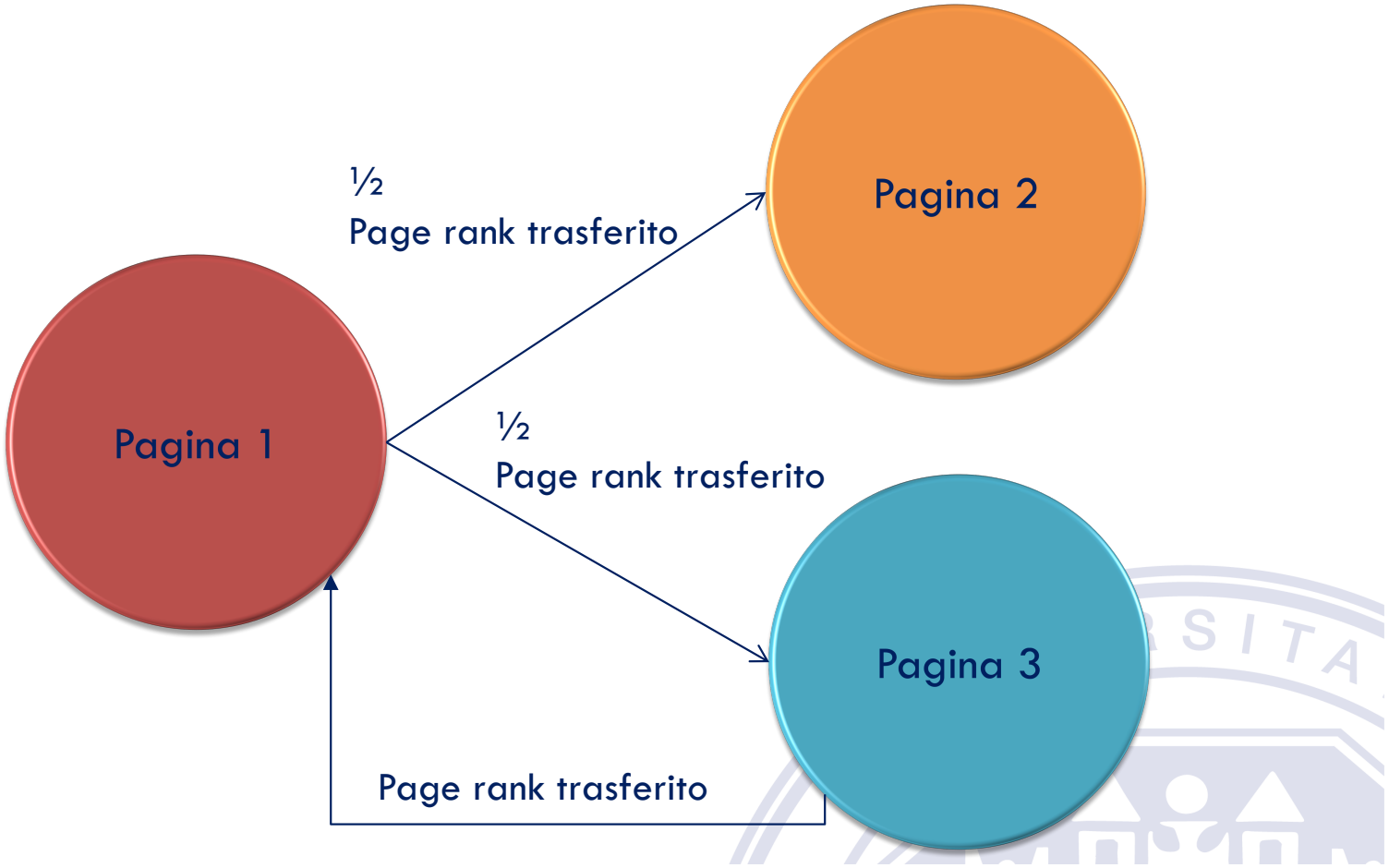
- Se una pagina ha rilevanza *X* e *n* link uscenti, **trasmette** una rilevanza *X/n* ai nodi della stella uscente
- **Rilevanza** di una pagina: **somma delle rilevanze ricevute** dai nodi della stella entrante

PageRank

PageRank


½
Page rank trasferito
Pagina 2
Pagina 1
½
Page rank trasferito
Pagina 3

PageRank

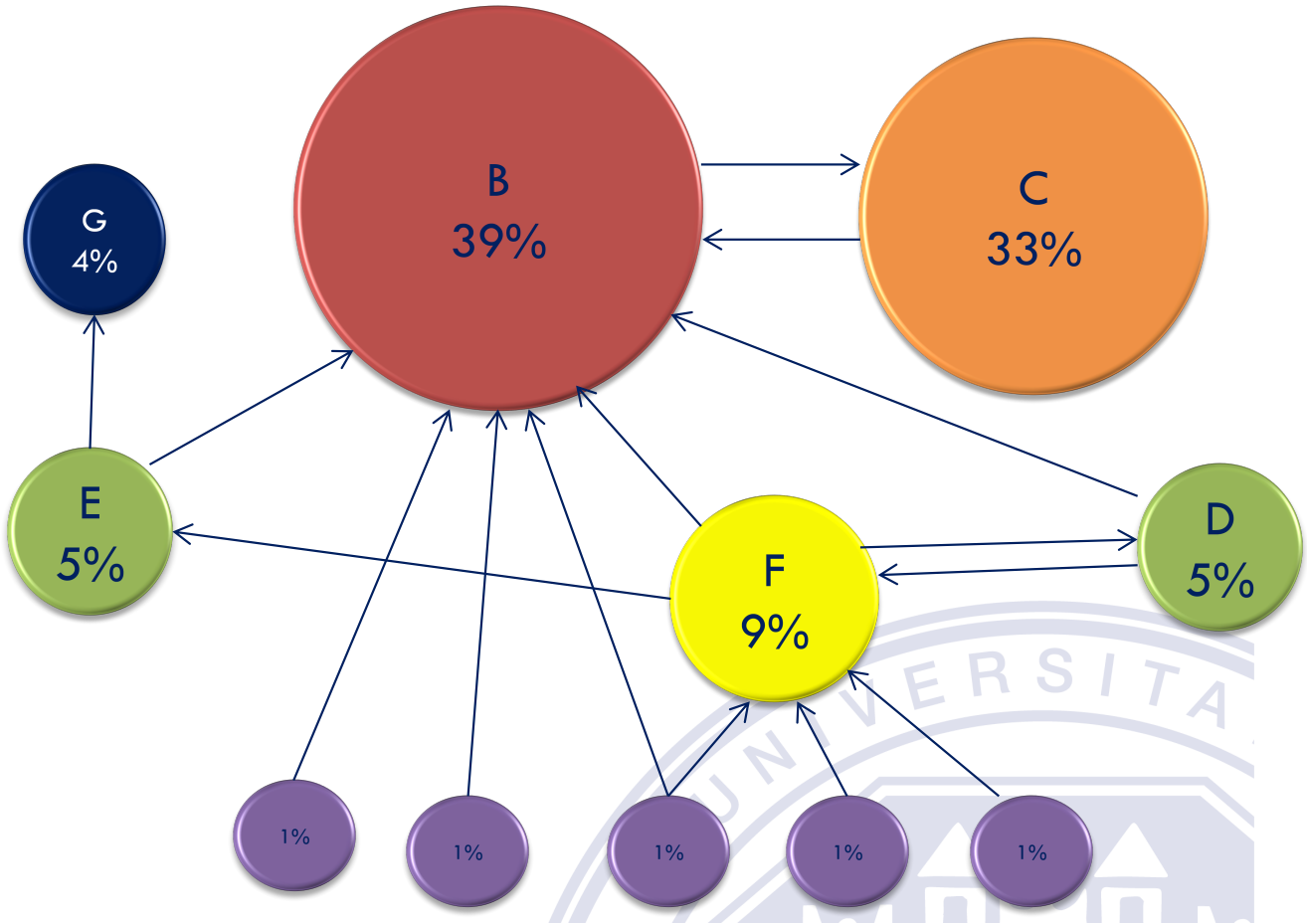

½
Page rank trasferito
Pagina 2
Pagina 1
½
Page rank trasferito
Pagina 3
Page rank trasferito

Motori di ricerca

Rilevanza di B trasferita solo a C

Rilevanza di C trasferita solo a B

Rilevanza di B: rilevanza trasferita da C, D, F, …



PageRank

- Calcolo *PageRank*: procedura ricorsiva **complessa**
- **Interpretazione**: **probabilità** che navigando in modo casuale nel Web **si visiti una determinata pagina**
- **Rilevanza** di una pagina **indipendente dalle parole chiave** utilizzate nelle ricerche
- Valore calcolabile **a priori**
  - Dipende solo dalla struttura del Web

Ricerca e Web

- Strategie per **condizionare** i motori di ricerca:
  - Creazione link verso una pagina per aumentarne l'autorevolezza
  - Vendita link
  - *Cloaking*: utilizzo versioni differenti del sito

SEO

- ***Seach Engine Optimization* (SEO)**: tecniche per migliorare il posizionamento nei motori di ricerca
- Obiettivo
  - **Migliorare la visibilità**, sia per motori di ricerca che per utenti
    - **Usabilità del sito**
    - **Link da altre pagine**

Tipi di ricerca

- Gli utenti effettuano ricerche per diversi scopi
- I principali:
  - **Navigazione**
  - **Informative**
  - **Transazionali**

SEO

On page

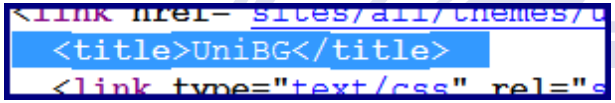
- Struttura
- Contenuto
- Aspetti tecnici

Off page

- Link
- Social media
- …

Costruzione della pagina

- Aspetti della singola pagina
  - **Titoli univoci e accurati** (parole chiave nel *tag title)*
  - I titoli sono mostrati come descrizione nella SERP



Costruzione della pagina

- Il m*eta* ***tag description*** potrebbe essere utilizzato l'informazione come ***snippet***
- In alternativa DMOZ Open Directory Project

La Repubblica.it - News in tempo reale - Le notizie e i video di politica ...
www.**repubblica**.it/
Repubblica è il quotidiano online aggiornato 24 ore su 24 su politica, cronaca, economia, sport, esteri, spettacoli, musica, cultura, scienza, tecnologia.

```
<meta name="description" content="Repubblica &egrave; il quotidiano online aggiornato 24 ore su 24 su politica,
```

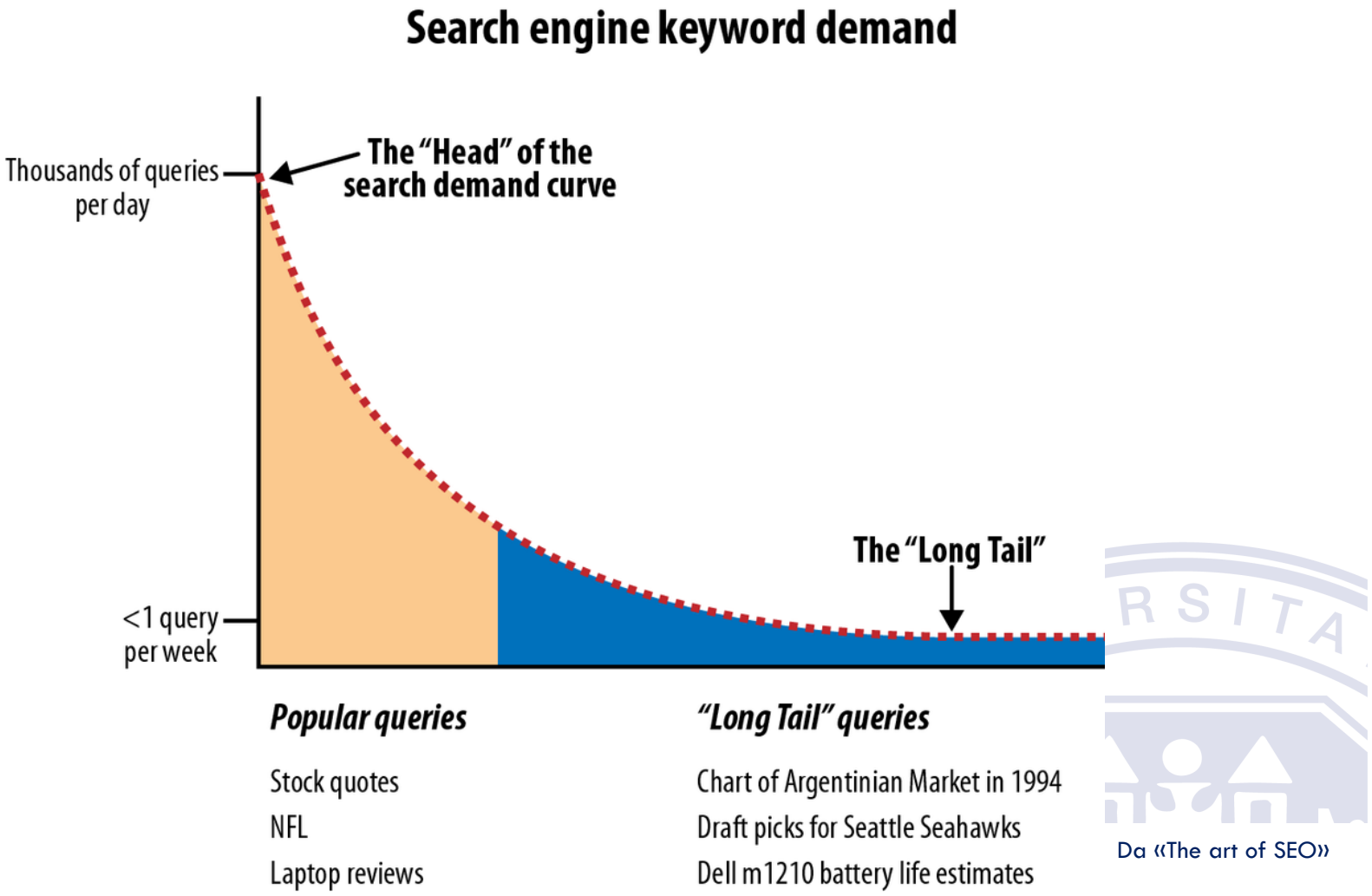
Navigazione

- La struttura di un sito è legata alla facilità di **navigazione**
- I motori di ricerca possono utilizzare la struttura per comprendere il **ruolo delle pagine**
- **Mappa del sito**
  - Mappa per gli utenti
  - Mappa per motori di ricerca (SiteMap XML)

SEO e contenuto

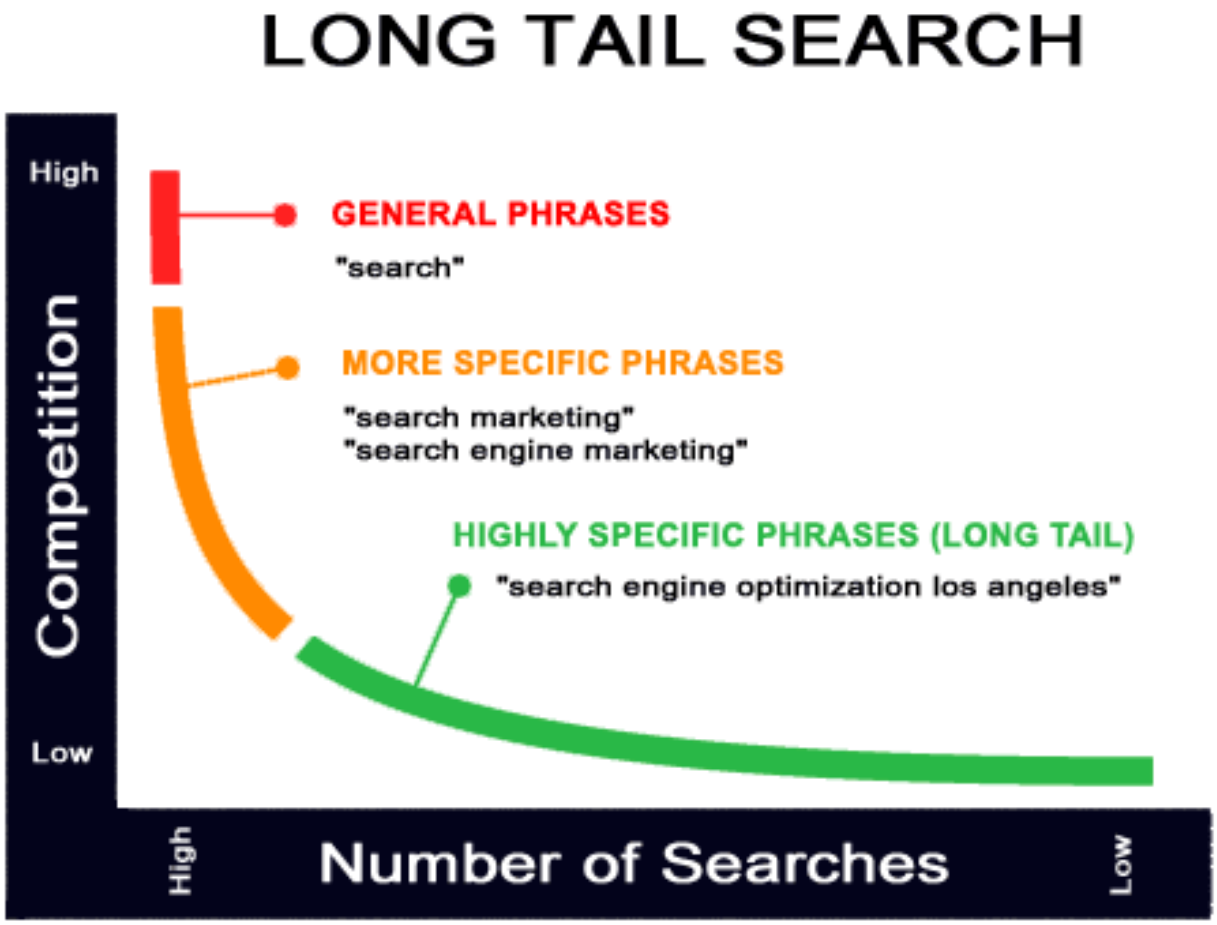
- Analisi delle **parole chiave**
  - Parole chiave maggiormente usate
  - Variazioni delle parole chiave
  - Relazione con la pubblicità (vedi GoogleAdWords)

Coda lunga delle ricerche



Da «The art of SEO»

Coda lunga delle ricerche



Da «The art of SEO»