Introduzione al crawling del WEB

Mauro Leoncini, Massimo Santini
Università di Modena e Reggio Emilia

Scaletta degli argomenti

- Che cos'è un crawler?
- Uno schema generale di crawler
- Uno sguardo ai problemi
- <parte del dott. Santini>

Che cos'è un web crawler?

- Un programma per la raccolta (scaricamento) di pagine web
  - Novità rispetto a IRS tradizionali
- Obiettivo: dato un insieme iniziale di url, scarica tutte le pagine raggiungibili dall'insieme seguendo gli iperlink
  - Tutte le pagine (crawler general purpose)
  - Solo quelle su determinati argomenti (focused crawler)

Quali pagine scaricare?

- File HTML (nodi o foglie dell'albero di visita) e file TEXT (foglie)
- Documenti "con struttura"
  - PDF
  - PS
  - PPT
  - DOC
  - ....

Schema di un crawler generico

- È dato un insieme di url iniziali u_1, u_2, ..., u_k
- Inserisci le u_i nel pool di url
- Inizializza il "repository" S delle pagine
- Finché ci sono url nel pool...
  - Estrai un url (u) dal pool
  - Scarica pagina di url u
  - Per ogni url v_i nella pagina scaricata
    - Inserisci v_i nel pool
  - Inserisci la pagina scaricata nel repository

Problemi (panoramica)

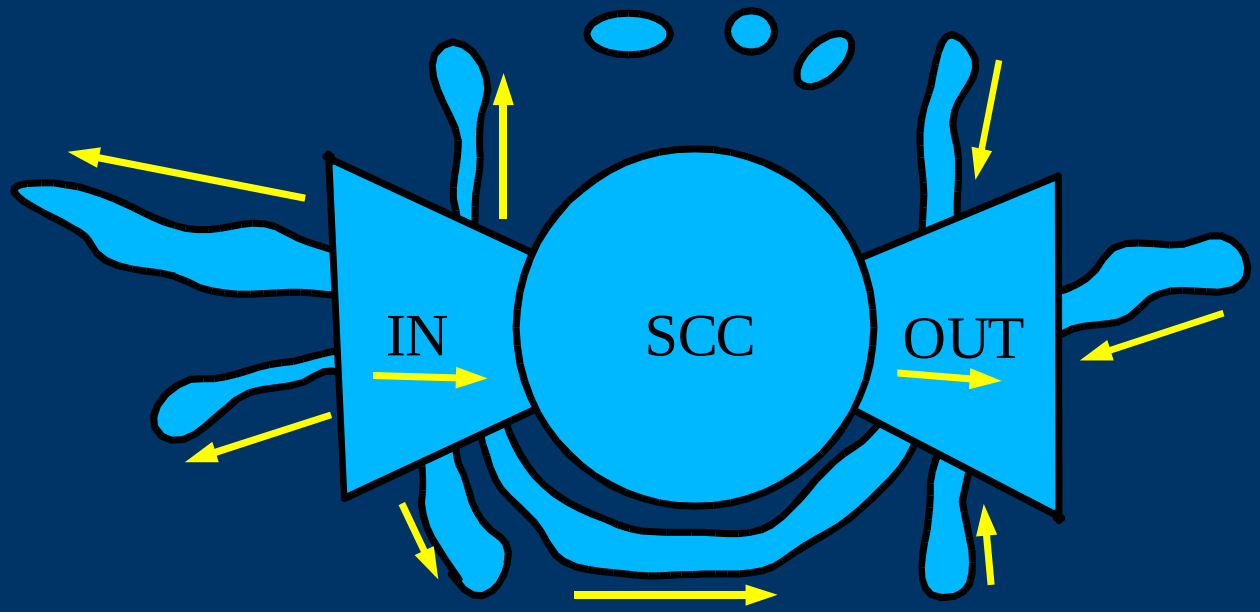
- Dipendono dalla scala (target dei crawl)
- Parsing delle pagine
- Eliminazione delle url già visitate
- Quale strategia di ricerca seguire (breadth- o depth-first search?)
- Politeness (come evitare di sovraccaricare un server o saturare la sua banda?)
- Fallimenti (ad esempio: time out, spider traps)
- Implementazione (es., come realizzare il repository e il pool?)
- Efficienza (come recuperare miliardi di pagine?)
  - Crawler paralleli

Eliminazione url già visitate

- Necessaria!
- Una possibilità: utilizzare una tabella hash (in cui le chiavi sono le url)
  - url -> fingerprint (hash value) di 32-128 bit
  - Piccoli crawl, nessun problema
  - Grandi crawl: 10^9 url, 64 bit fingerprint => 8GB
- Accesso su disco lento e non si sfrutta la località
- Una possibile soluzione: hash a due livelli
  - http://www.dsi.univp.it/people/index.html
  - H_1(www.dsi.univp.it) -> 24 bits
  - H_2(/people/index.html) -> 40 bits
  - Utilizzare un B-tree con chiave H_1H_2

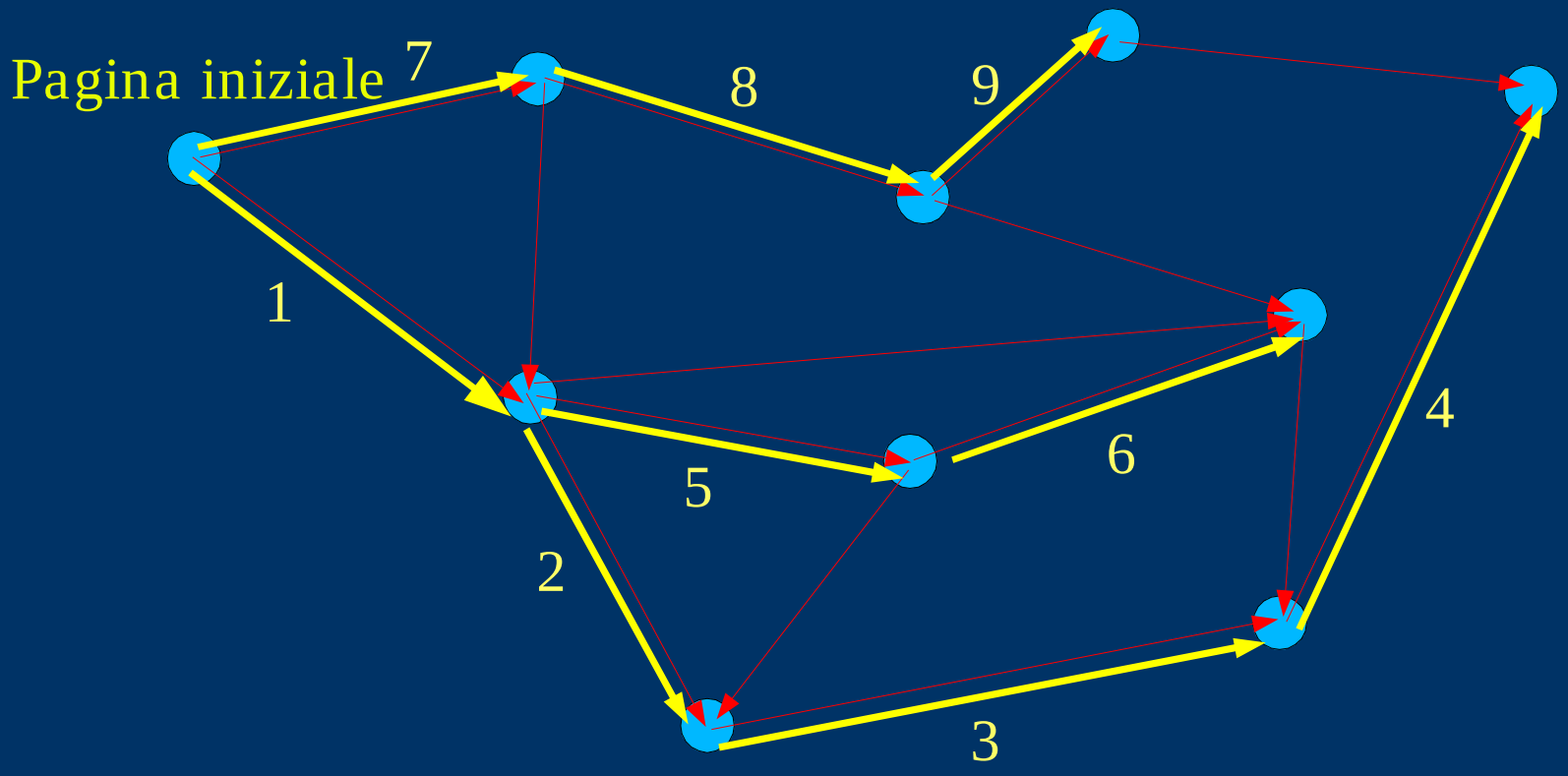
Esplorazione

- Struttura del web (bowtie)

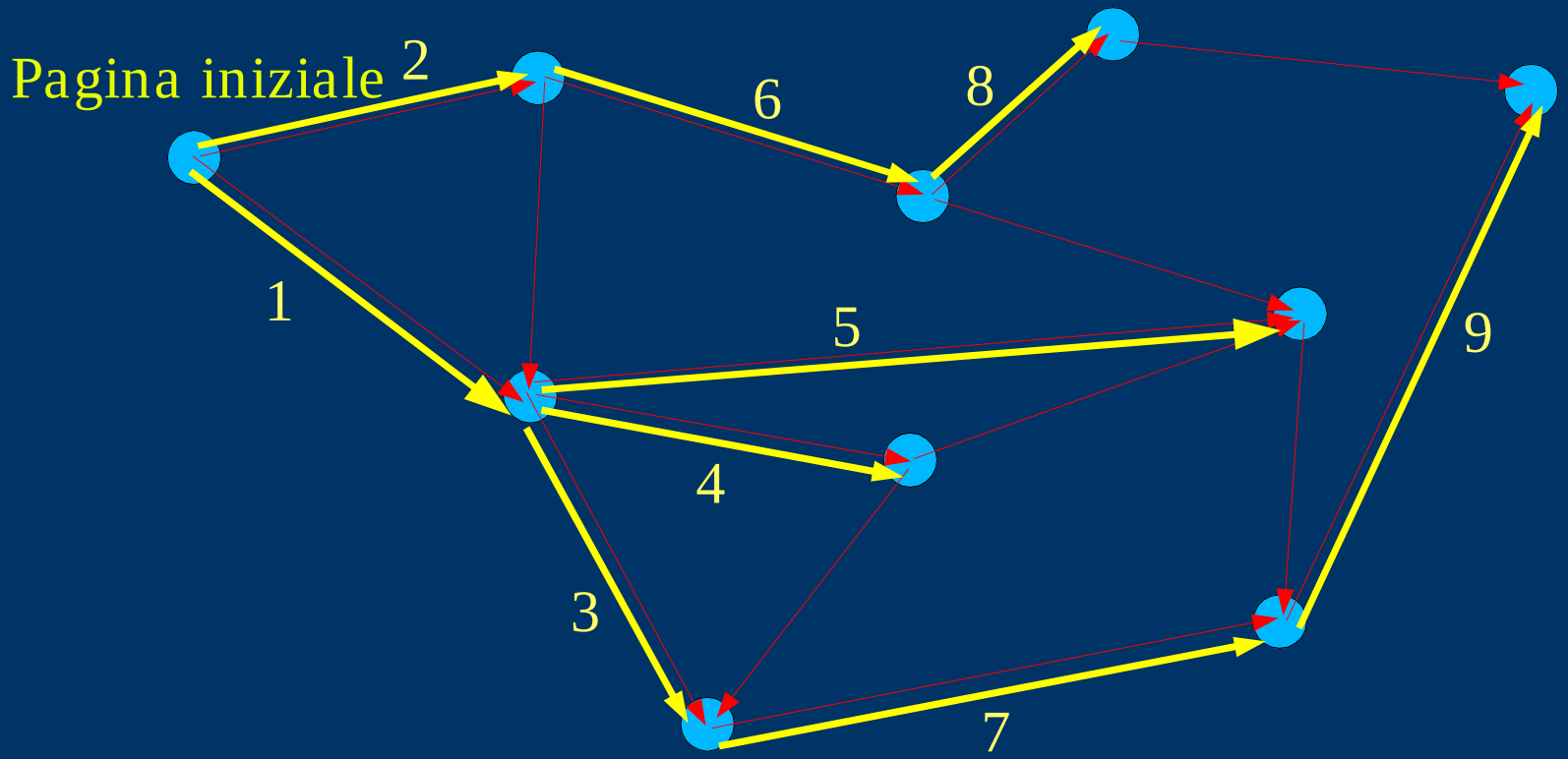


- Evoluzione temporale
- Pagine dinamiche
- => Nessuna speranza di coprire tutto il web

Depth-first search



Breadth-first search



Un crawler in un comando

- wget -r -l depth -H -w seconds url
  - Reperimento ricorsivo (opzione -r)
  - Fino ad una distanza massima dall'url iniziale pari a depth link (opzione -l depth)
  - Attraversando host diversi (opzione -H)
  - Attendendo seconds secondi prima di ogni tentativo (opzione -w seconds)
- Esempio (se abbiamo connettività in aula ...)