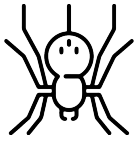
ALGORITMICA PER IL WEB

Prof. Sebastiano Vigna
6 CFU

Luca Cappelletti

Lecture Notes
Year 2017/2018



Magistrale Informatica
Università di Milano
Italy
10 ottobre 2018

Indice

1 Introduzione

Definizione 1.0.1 (Path (Cammino)). Una sequenza di **vertici** $C = \{x_0, \ldots, x_n\}$ con $x_i \in V$, tale che:

$$\forall x_i, x_{i+1} \in C \quad x_i \rightarrow x_{i+1}$$

Definizione 1.0.2 (Vertice coraggiungibile). Un vertice x è corraggiungibile da un vertice y se invertendo le frecce risulta raggiungibile da y.

Definizione 1.0.3 (Grado di un vertice). Il grado positivo di un vertice è dato dal numero di **successori** mentre il grado negativo è dato dal numero di **predecessori**.

2 Crawling

2.1 Politeleness (buona educazione)

Si riferisce, nel crawling, ad un insieme di strategie utilizzate per evitare di essere bloccati da siti e provider.

Questa deve essere vincolata per **host** e per **ip**, e consiste nell'evitare di sovraccaricare la macchina che offre il servizio per esempio facendo una pausa tra una richiesta e quella successiva, magari proporzionale al tempo di scaricamento.

È altamente suggerito cercare di seguire le regole di crawling che ogni sito esplicita nel proprio robots.txt.

2.1.1 Struttura dati degli url nel crawler: coda con ritardo

La struttura è composta da una **coda di siti con priorità** e per ogni sito una **coda di url**. Ogni sito avrà associato un timestamp, rappresentante il tempo dopo il quale si potrà nuovamente fare richieste al sito rispettando la politeleness.

Comprendendo anche gli **IP** nella coda, si procede ad aggiungere un nuovo primo livello alla coda, che ora diviene una coda di ip con associate code di host a cui a loro volta vengono associate code di url.