Università degli Studi di Catania
Dipartimento di Matematica e Informatica
Corso di laurea in Informatica Magistrale
Compilatori anno 2014-15

Web crawler Phoneutria

Introduzione

Il presente documento riguarda la descrizione del progetto “phoneutria” , un web crawler scritto in linguaggio C, a cui vengono dati in input uno o più semi iniziali (URL) e delle keywords di ricerca.

Per ogni URL trovato (con eventuali limitazioni), partendo da quello iniziale, viene ricercata la presenza delle keywords data in input all’interno della pagina.

Struttura del progetto

Fase di download

La fase di download si occupa di scaricare tutte le pagine che sono presenti all’interno di un set di URL (aggiornata di volta in volta aggiungendo gli URL trovati nelle diverse pagine, con limitazioni riguardanti la profondità degli URL stessi).
In dettaglio si inizializza il set di URL con il/i seed iniziale/i dato/i in input. Successivamente per ogni URL del set vengono eseguite le seguenti operazioni:

1. Si cerca di risolvere l’URL;
2. Se la fase 1 ha successo viene aperta una socket sul precedente URL;
3. Tramite la socket viene inviata una richiesta http che permette di scaricare la pagina (o il file) puntato dall’URL;
4. La pagina (o il file) ottenuta viene dato in input alla funzione di parsing.

Fase di parsing

La fase di parsing si occupa di scansionare le pagine ottenute dalla fase precedente e ricercare al loro interno le keywords prese in input e l'esistenza di altri URL.
In dettaglio la pagina viene scansionata carattere per carattere e:

- Se viene trovato un url, tramite il calcolo di una funzione hash viene controllato se l'url era già stato inserito nel set. Se non è presente, allora viene inserito all'interno del set(si vedano specifiche del set sotto);
- Se viene trovata la keyword, la pagina viene memorizata all'interno del file system.

Le estensioni delle pagine e dei file accettati dal parser (quindi aggiunti nel set di url) sono:

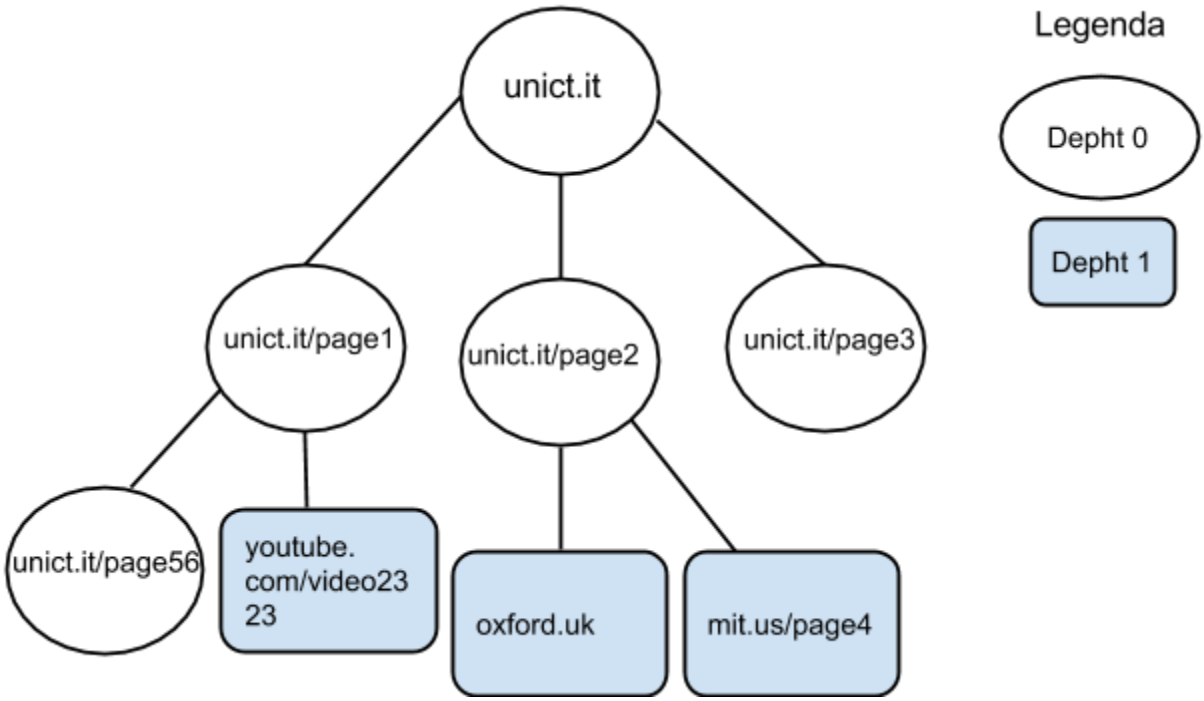
- html;
- htm;
- xhtml;
- xml;
- php;
- txt;
- asp;
- aspx;
- jsp;
- jspx;
- do.

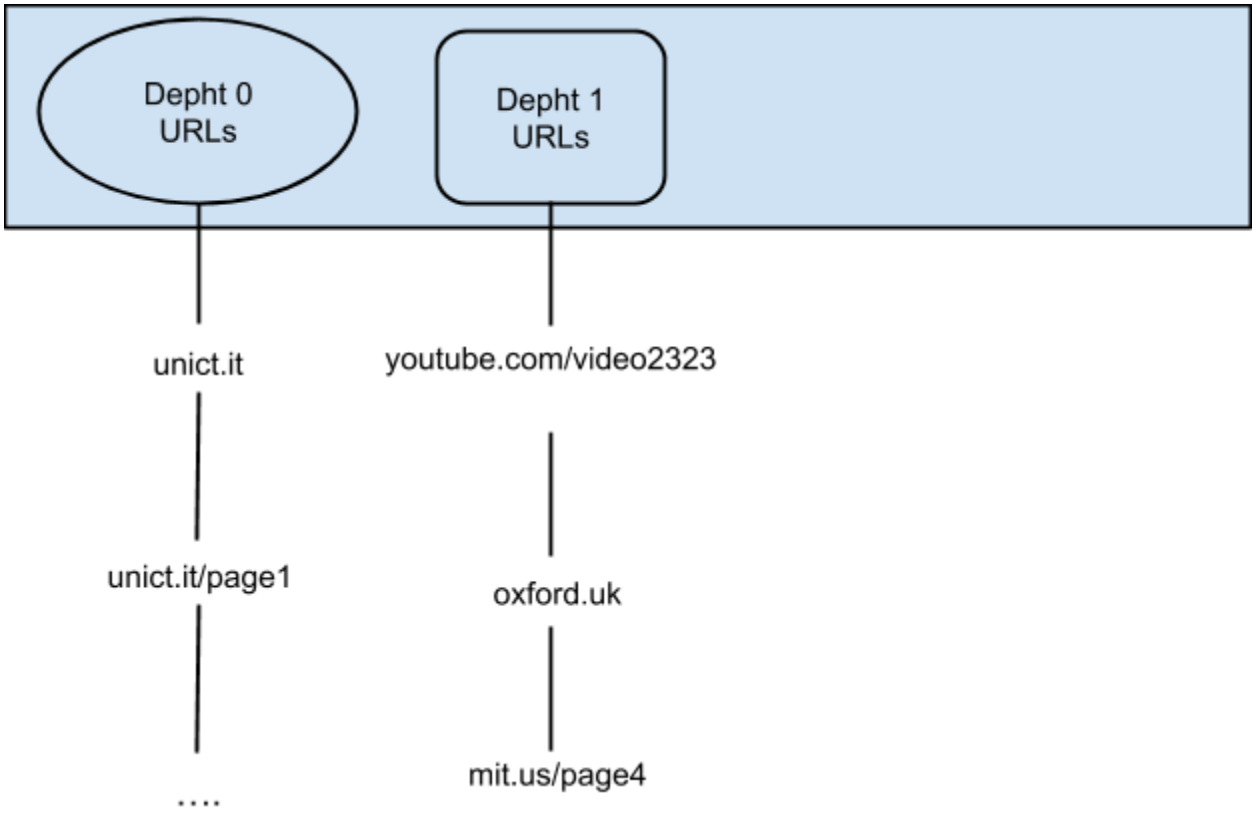
Set di URL

Il set di URL è una struttura dati così formata:

- ogni dominio trovato viene inserito in un nuovo nodo aggiunto all'interno di una coda (lista linkata);
- ogni sottodominio trovato viene aggiunto in uno stack puntato dal nodo riferito al suo dominio.

Il primo nodo della coda contiene lo stack riferito al dominio del seed iniziale che viene analizzato totalmente. Invece gli stack degli altri nodi vengono analizzati solo alla profondità settata, in modo da limitare le ricerche.


unict.it
unict.it/page1
unict.it/page2
unict.it/page3
unict.it/page56
youtube.
com/video23
23
oxford.uk
mit.us/page4
Legenda
Depht 0
Depht 1



Credits

Have participated at project (Alphabetical Order):

- Cantarella Danilo (http://cantarelladanilo.com)
- Maccarrone Roberta (http://robertamaccarrone.altervista.org)
- Parasiliti Parracello Cristina (http://parasiliticristina.altervista.org)
- Randazzo Filippo (http://randazzofilippo.com)
- Ramon Gago (http://ramongagocarrera.wix.com/spain)
- Safarally Dario (http://dariosafarally.altervista.org)
- Siragusa Sebastiano (http://sebastianosiragusa.altervista.org/)
- Vindigni Federico (http://federicovindigni.altervista.org)