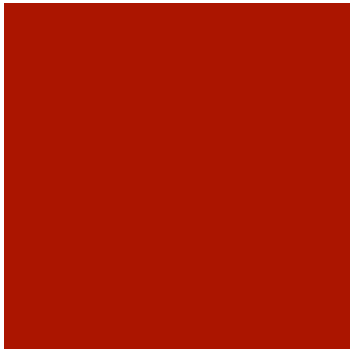
SAPIENZA
UNIVERSITÀ DI ROMA

Giovanni Stilo, Ph.D.
stilo@di.uniroma1.it

Chasing the White Rabbit

Simple Crawling principles

Requirements



- Curiosity!
- Network Principles
  - TCP/IP protocol
  - HTTP protocol
  - DNS
- W.W.W. Principles
  - Html language
  - JavaScript language
- IR background
- UML (highly recommended)
- Design Pattern (highly recommended)
- Maven 2 (Optional)

Web *

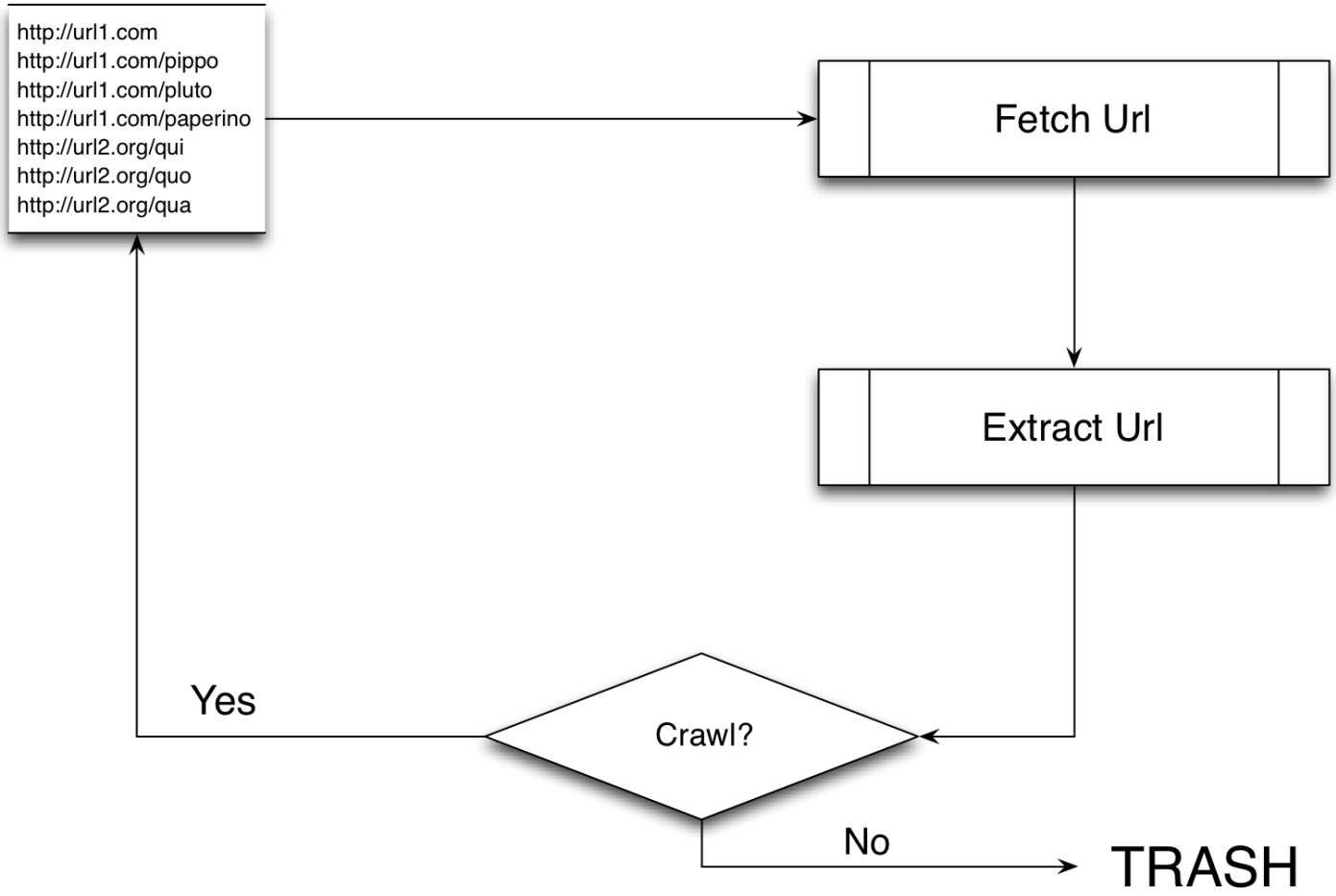
- A **Web crawler** is an Internet bot that systematically *browses* the World Wide Web, typically for the purpose of Web indexing.
- **Web indexing** refers to various methods for indexing the contents of a website or of the Internet as a whole. Web indexing is also becoming important for periodical websites.
- **Web scraping** (web harvesting or web data extraction) is a computer software technique of extracting information from websites. Such software programs simulate human exploration of the World Wide Web.

Web Crawler

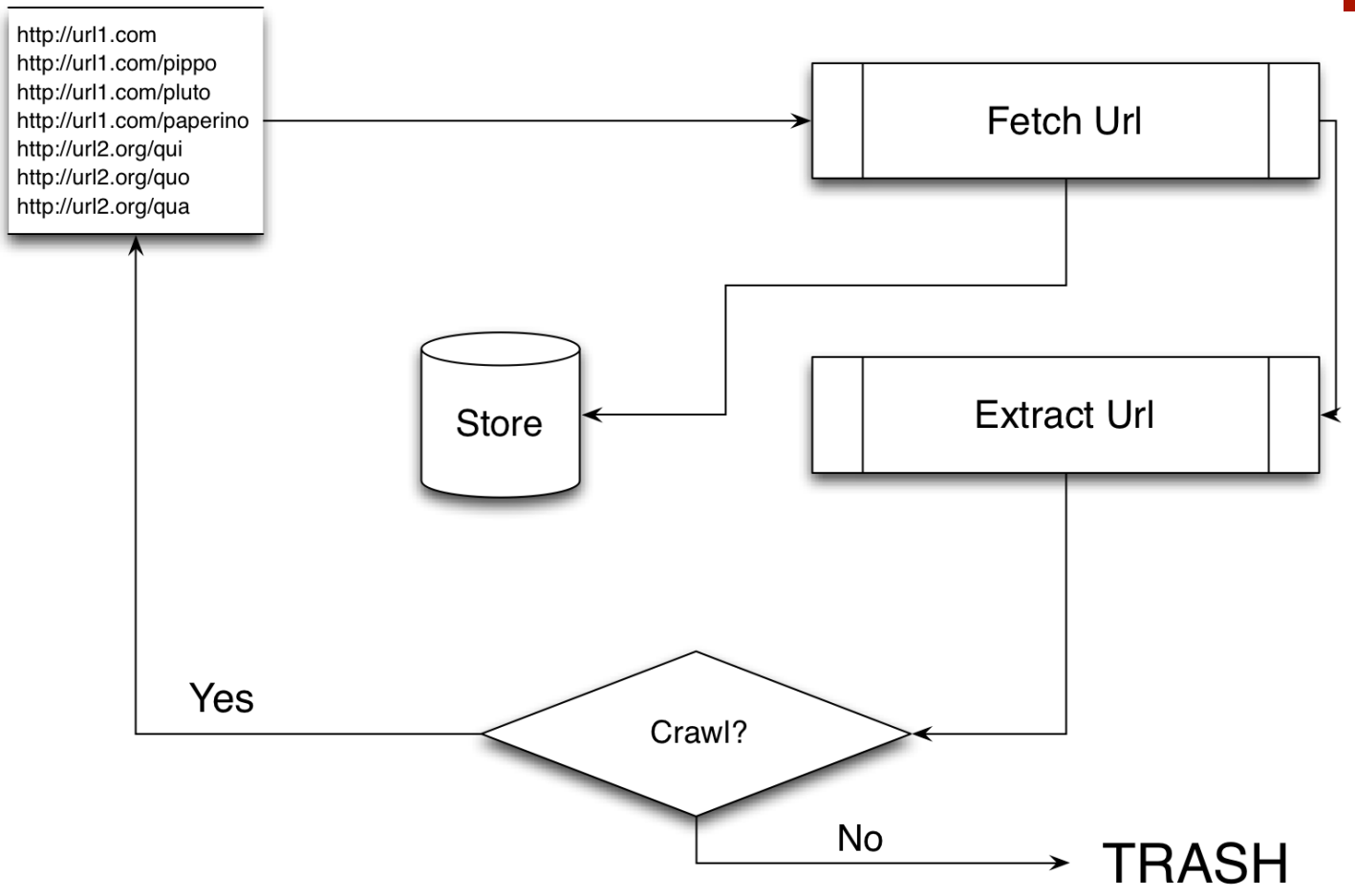
- A Web crawler starts with a list of URLs to visit, called the ***seeds***.
- As the crawler visits these URLs, it identifies all the hyperlinks in the page and adds them to the list of URLs to visit, called the ***crawl frontier***.
- URLs from the frontier are recursively visited according to a set of ***policies***.

- **GREAT! LOOKS SIMPLE!**

Simple Crawler



Crawler


http://url1.com
http://url1.com/pippo
http://url1.com/pluto
http://url1.com/paperino
http://url2.org/qui
http://url2.org/quo
http://url2.org/qua
Fetch Url
Store
Extract Url
Yes
Crawl?
No
TRASH

Facts

- **634 million –** Number of websites (December).
- **51 million –** Number of websites added during the year.
- **87.8 million –** Number of Tumblr blogs..
- **59.4 million –** Number of WordPress sites around the world.
- **35% –** The average web page became this much larger during 2012.
- **4% –** The average web page became this much slower to load during 2012.
- **~45 billion** pages indexed by Google

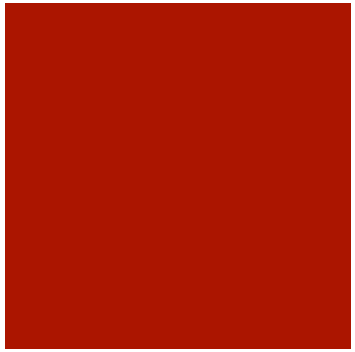
Issues

- The large volume implies that the crawler can only download a limited number of the Web pages within a given time, so it needs to prioritize its downloads. The high rate of change implies that the pages might have already been updated or even deleted.
- Edwards *et al.* noted, "Given that the bandwidth for conducting crawls is neither ***infinite nor free***, it is becoming **essential** to crawl the Web in not only a scalable, but efficient way, if some reasonable measure of quality or freshness is to be maintained."
- A crawler **must carefully** choose and how at each step which pages to visit next.

Related

- A crawler must not only have a **good crawling strategy**, as noted in the previous sections, but it should also have a **highly optimized architecture**.
- While it is *fairly easy to build a slow crawler* that downloads a few pages per second for a short period of time, building a high-performance system that can download hundreds of millions of pages over several weeks presents **a number of challenges** in system design, I/O and network efficiency, and robustness and manageability.

Excercise



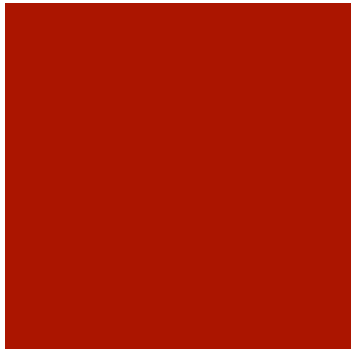
Realize the **Design** of a Crawler

Be abstarct as much as possible considering all future possible operation.

For example you need:

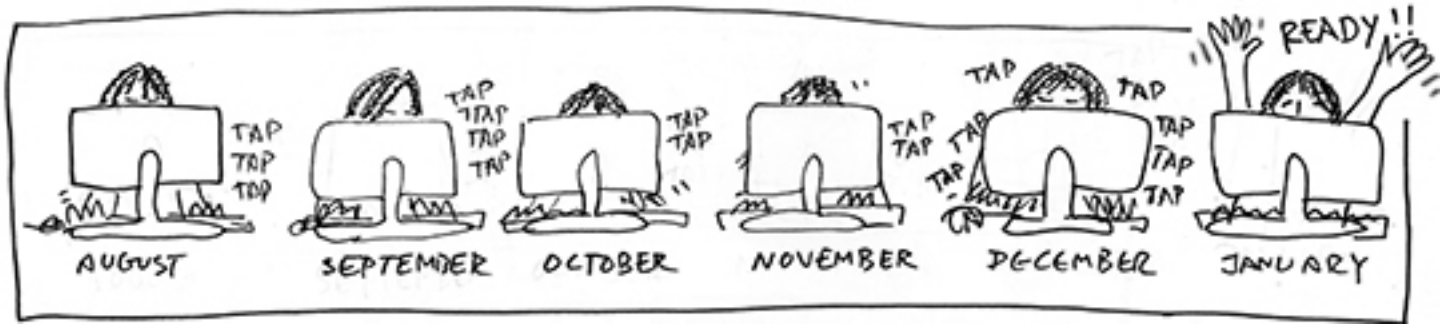
- **Url Manager ?**
- **Fetcher ?**
- **Action Manager ?**
- **… ?**
- **… ?**
- **… ?**

Users Crawling



- It is possible to work in the same way of a crawler work over every network/graph.
- For example we should crawling users of a Social Network. How?
  - seeds = list of starting user to crawl
  - extracted url = list of friends/follower
  - At each step you crawl a new user and relative information like normally get a web-page.
- It is possible to find the problems during the users crawling. So should be useful:
  - Maintainng a list/map of visited users
  - Create a queue of users to download
  - Perform some kind of operation over downloaded data (store…).

Let's Try?!?!


TAP
TAP
TAP
AUGUST
TAP
TAP
TAP
TAP
SEPTEMBER
TAP
TAP
OCTOBER
TAP
TAP
NOVEMBER
TAP
TAP
TAP
TAP
TAP
TAP
TAP
DECEMBER
READY!!
JANUARY