**WEB CRAWLER**

Crawling the web has been an area of research and intersrt for long just because of the intrinsic value of the target data .The document retrieved in web search have been found to truthful in majority of cases .This has led to study for better understanding of the structure of web and the layout of documents so that more efficient algorithms can be developer for crawling

Utility of web crawlers

Social media

Concept of social media has come from the need of humans to connect with people of common interests. e.g keeping up with friends from school,college or workplaces

Social networking has been the studty of social sciences,computer sciences,statistics and economics long before the advent of internet social media platforms

Applications:

Epidemology,Espionage

Architecture of crawler

It is the process of capturing and extracting data from social networks, such as Facebook, Twitter and LinkedIn. This data let you sense consumer behavior, trends and sentiments. Eventually, you get the pan insight, be it of your customers or retailers or competitors, for carrying out business research.

To understand how social media [data scraping](http://www.multitechit.com.au/web-scraping-services.php) takes place, you should know that it runs on a piece of code. It is called scraper. As it runs, the “Get” query rolls out to extract the HTML data coming from Facebook or any other social channels.

Thereafter, algorithms analyse a string of symbols, either in natural language or computer language or models in the Document Object Model (DOM) structure. This parsing process determines nodes (an object representing a part of the document).  Then, it creates a node processor to show output in a normalized format. In simple words, the scraper comes into play, filtering through the data to pick up the requisite data sets. Once the requirement is fulfilled, the data is translated into a specific format.

In the nutshell, a code is tailored to:

**a.** Recognise unique HTML site structures

**b.**Extract  and transform data

**c.**Store the captured data

**d.**Extract data from APIs

Challenges

1. Blocked
2. Scraping refused

Solutions

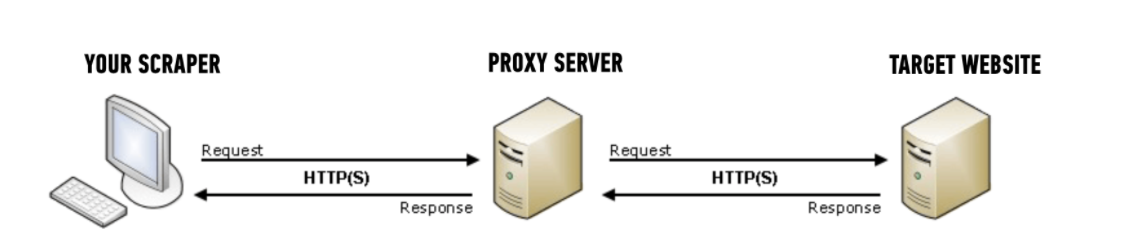
1. Getting blocked

Most popular websites have software to detect huge number of requests from one IP address

and blocking the ip suspecting DDOS attack

Solution:

Using Proxy server



Benefits of proxy server

Here are two main benefits to using proxies for your web scraping project:

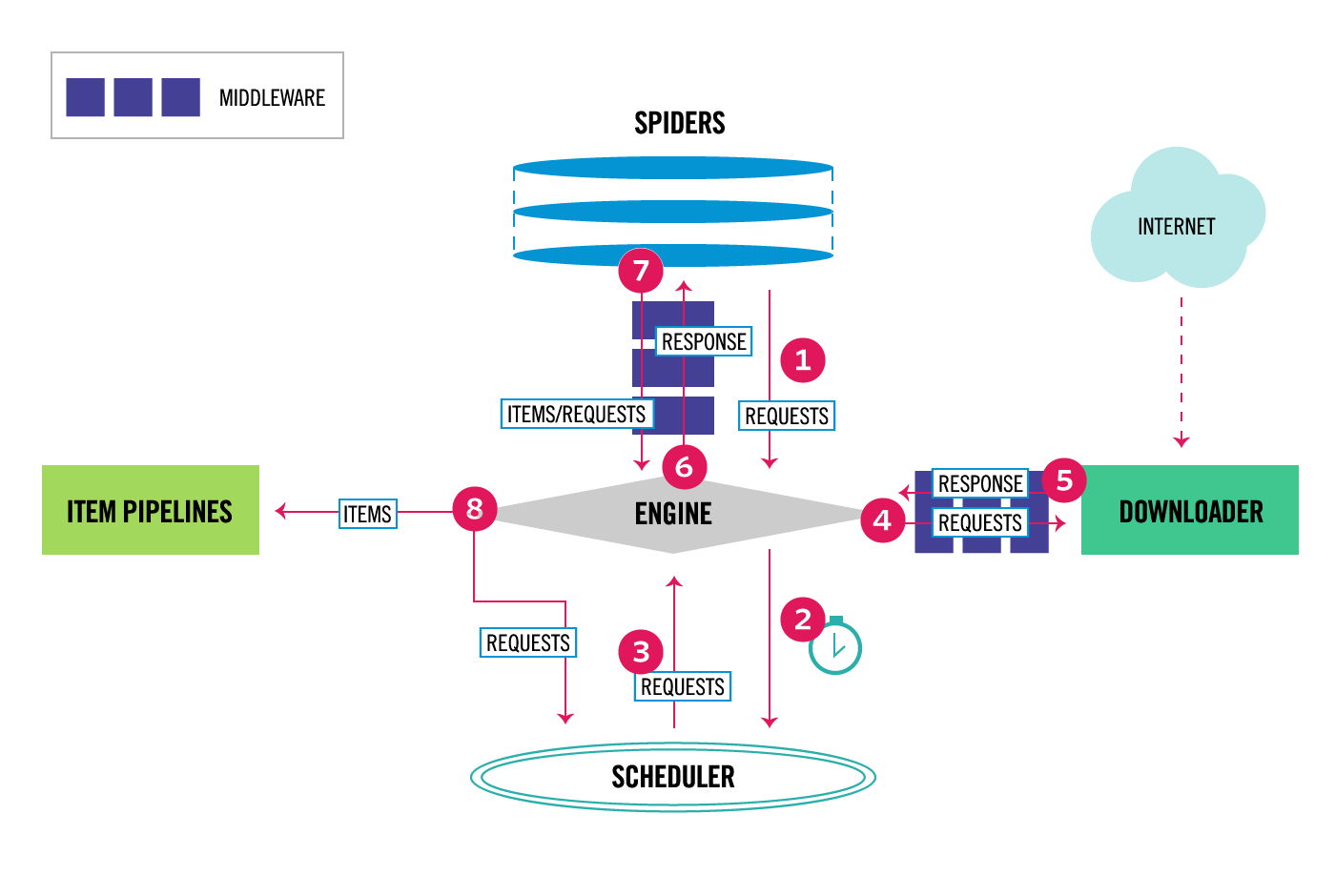
1. Hiding your source machine’s IP address
2. Getting past rate limits on the target site

Many large sites have software in place to detect when there are a suspicious number of requests coming in from one IP address, since this usually indicates some sort of automated access – it could be scraping, or something security related like fuzzing.

The way this rate limiting software is usually setup, if too many requests come in from one IP address in a short amount of time, then the site will return some sort of error message to “block” future requests from that client for a pre-set period of time.

Techniques to overcome anti-scraping mechanism

* proxy rotation
* faking user-agents
* Insering waits between request



References

* Social network theory (cf. Wasserman & Faust)