Captation de données web

Camille Maussang

camille.maussang@linkfluence.net

IC05 - P10

Qui suis-je ?

Qui suis-je ?

- Camille Maussang

Qui suis-je ?

- Camille Maussang
- Responsable de l'équipe ingénierie chez linkfluence...

Qui suis-je ?

- Camille Maussang
- Responsable de l'équipe ingénierie chez linkfluence...
- ... qui développe des outils d'analyse du web social

Qui suis-je ?

- Camille Maussang
- Responsable de l'équipe ingénierie chez linkfluence...
- ... qui développe des outils d'analyse du web social
- ... en captant des données sur le web ;)

Qu'est-ce que le web ? Définition historique

Le web est un corpus de documents

Qu'est-ce que le web ? Définition historique

Le web est un corpus de documents

- ouvert,

Qu'est-ce que le web ? Définition historique

Le web est un corpus de documents

- ouvert,
- hétérogène,

Qu'est-ce que le web ? Définition historique

Le web est un corpus de documents

- ouvert,
- hétérogène,
- et dynamique.

Qu'est-ce que le web ? Définition historique

Le web est un corpus de documents

- ouvert,
- hétérogène,
- et dynamique.

En résumé

Qu'est-ce que le web ? Définition historique

Le web est un corpus de documents

- ouvert,
- hétérogène,
- et dynamique.

En résumé

Le web c'est

Qu'est-ce que le web ? Définition historique

Le web est un corpus de documents

- ouvert,
- hétérogène,
- et dynamique.

En résumé

Le web c'est n'importe qui, n'importe où (ouvert)

Qu'est-ce que le web ? Définition historique

Le web est un corpus de documents

- ouvert,
- hétérogène,
- et dynamique.

En résumé

Le web c'est n'importe qui, n'importe où (ouvert) qui publie n'importe quoi (hétérogène)

Qu'est-ce que le web ? Définition historique

Le web est un corpus de documents

- ouvert,
- hétérogène,
- et dynamique.

En résumé

Le web c'est n'importe qui, n'importe où (ouvert) qui publie n'importe quoi (hétérogène) n'importe quand (dynamique).

Qu'est-ce que le web ?

Au départ, le web est un corpus de documents

Qu'est-ce que le web ?

Au départ, le web est un corpus de documents

- Un langage de description de documents : `HTML`

Qu'est-ce que le web ?

Au départ, le web est un corpus de documents

- Un langage de description de documents : `HTML`
- Un protocole d'adressage : `URL`

Qu'est-ce que le web ?

Au départ, le web est un corpus de documents

- Un langage de description de documents : `HTML`
- Un protocole d'adressage : `URL`
- Un protocole de transport : `HTTP`

Qu'est-ce que le web ?

Au départ, le web est un corpus de documents

- Un langage de description de documents : `HTML`
- Un protocole d'adressage : `URL`
- Un protocole de transport : `HTTP`

Peu structuré

Qu'est-ce que le web ?

Au départ, le web est un corpus de documents

- Un langage de description de documents : `HTML`
- Un protocole d'adressage : `URL`
- Un protocole de transport : `HTTP`

Peu structuré

- Pas de normes mais des recommandations

Qu'est-ce que le web ?

Au départ, le web est un corpus de documents

- Un langage de description de documents : `HTML`
- Un protocole d'adressage : `URL`
- Un protocole de transport : `HTTP`

Peu structuré

- Pas de normes mais des recommandations
- Des standards de facto

Qu'est-ce que le web ?

Au départ, le web est un corpus de documents

- Un langage de description de documents : `HTML`
- Un protocole d'adressage : `URL`
- Un protocole de transport : `HTTP`

Peu structuré

- Pas de normes mais des recommandations
- Des standards de facto
- Liberté au publieur de faire ce qu'il veut

Qu'est-ce que le web ?

Aujourd'hui, le web devient un corpus de ressources

Qu'est-ce que le web ?

Aujourd'hui, le web devient un corpus de ressources

- Document (article, commentaire, photo, vidéo, statut)

Qu'est-ce que le web ?

Aujourd'hui, le web devient un corpus de ressources

- Document (article, commentaire, photo, vidéo, statut)
- Utilisateur (profil, ami, *follower*)

Qu'est-ce que le web ?

Aujourd'hui, le web devient un corpus de ressources

- Document (article, commentaire, photo, vidéo, statut)
- Utilisateur (profil, ami, *follower*)
- Application

Comment saisir le web ?

Le web peut être représenté par des graphes

Comment saisir le web ?

Le web peut être représenté par des graphes

- où les noeuds sont :

Comment saisir le web ?

Le web peut être représenté par des graphes

- où les noeuds sont :
  - des pages,

Comment saisir le web ?

Le web peut être représenté par des graphes

- où les noeuds sont :
  - des pages,
  - des sites,

Comment saisir le web ?

Le web peut être représenté par des graphes

- où les noeuds sont :
  - des pages,
  - des sites,
  - des mots,

Comment saisir le web ?

Le web peut être représenté par des graphes

- où les noeuds sont :
  - des pages,
  - des sites,
  - des mots,
  - des profils,

Comment saisir le web ?

Le web peut être représenté par des graphes

- où les noeuds sont :
  - des pages,
  - des sites,
  - des mots,
  - des profils,
  - des *ressources*,

Comment saisir le web ?

Le web peut être représenté par des graphes

- où les noeuds sont :
  - des pages,
  - des sites,
  - des mots,
  - des profils,
  - des *ressources*,
- et les arcs des liens.

Comment saisir le web ?

Capter des données sur le web requiert un certain nombre de ressources (bande passante, stockage, temps machine, etc.) :

Comment saisir le web ?

Capter des données sur le web requiert un certain nombre de ressources (bande passante, stockage, temps machine, etc.) :

- Que cherchons-nous ?

Comment saisir le web ?

Capter des données sur le web requiert un certain nombre de ressources (bande passante, stockage, temps machine, etc.) :

- Que cherchons-nous ?
- Que faire pour récupérer ce qui nous est important ?

Comment saisir le web ?

Capter des données sur le web requiert un certain nombre de ressources (bande passante, stockage, temps machine, etc.) :

- Que cherchons-nous ?
- Que faire pour récupérer ce qui nous est important ?
- Toujours penser « heuristiques »...

Comment saisir le web ?

Capter des données sur le web requiert un certain nombre de ressources (bande passante, stockage, temps machine, etc.) :

- Que cherchons-nous ?
- Que faire pour récupérer ce qui nous est important ?
- Toujours penser « heuristiques »...
- ... et « effets de bord »!

Carte de sites

Crawler

Principe

Crawler

Principe

- Télécharger 1 page

Crawler

Principe

- Télécharger 1 page
- Extraire les liens

Crawler

Principe

- Télécharger 1 page
- Extraire les liens
- Télécharger les pages pointées par les liens

Crawler

Principe

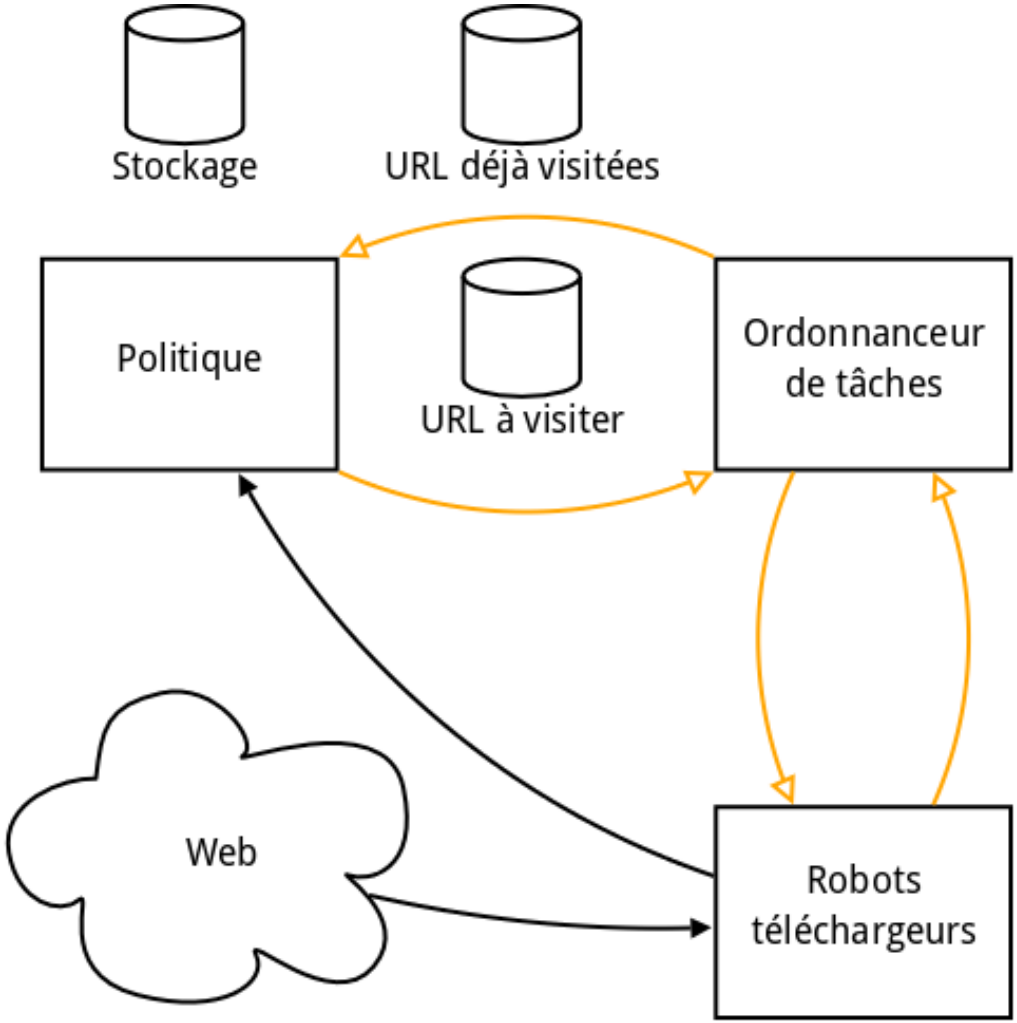
- Télécharger 1 page
- Extraire les liens
- Télécharger les pages pointées par les liens
- etc. etc.

Crawler - Exemple

```
1   use strict; use warnings;
2   use LWP::Simple;
3
4   my ( $max_depth, @seed ) = @ARGV or die( 'need depth and url(s)' );
5   my @already_visited = ();
6   my $depth = 0;
7   my @to_visit = @seed;
8
9   while( $depth <= $max_depth && @to_visit ) {
10      print "crawling depth $depth\n";
11      my @links = ();
12      for my $url ( @to_visit ) {
13          if( my $content = get( $url ) ) {
14              while ( $content =~ m/<a href="([^"]+)"/gi) { push @links, $1 }
15          }
16          push @already_visited, $url;
17          print "$url visited.\n";
18      }
19      @to_visit = ();
20      for my $url_to_check ( @links ) {
21          my $to_push = 0;
22          for my $url_visited ( @already_visited ) {
23              if( $url_to_check eq $url_visited ) { $to_push = 0; last; }
24              $to_push = 1;
25          }
26          push @to_visit, $url_to_check
27              if( $to_push && !grep{ $_ eq $url_to_check } @to_visit );
28      }
29      $depth++;
30  }
31  print "end.\n";
```

linkfluence

Crawler - Architecture



Crawler - Extraction

Principe du scraping

Crawler - Extraction

Principe du scraping

Analyser une page web pour en extraire une information spécifique

Crawler - Extraction

Principe du scraping

Analyser une page web pour en extraire une information spécifique

Problèmes

Crawler - Extraction

Principe du scraping

Analyser une page web pour en extraire une information spécifique

Problèmes

- Validité du code `HTML`

Crawler - Extraction

Principe du scraping

Analyser une page web pour en extraire une information spécifique

Problèmes

- Validité du code `HTML`
- Encodage

Crawler - Extraction

Principe du scraping

Analyser une page web pour en extraire une information spécifique

Problèmes

- Validité du code `HTML`
- Encodage
- `DOM` ou Regexp ou les deux

Crawler - Extraction

Principe du scraping

Analyser une page web pour en extraire une information spécifique

Problèmes

- Validité du code `HTML`
- Encodage
- `DOM` ou Regexp ou les deux
- Template et dynamisme des pages scrapées

Crawler - Extraction

Principe du scraping

Analyser une page web pour en extraire une information spécifique

Problèmes

- Validité du code `HTML`
- Encodage
- `DOM` ou Regexp ou les deux
- Template et dynamisme des pages scrapées
- Flash et Javascript

Crawler - Difficultés

- Adressage

Crawler - Difficultés

- Adressage
  - Normalisation d'URL (doublons)

Crawler - Difficultés

- Adressage
  - Normalisation d'URL (doublons)
  - Site ou page ?

Crawler - Difficultés

- Adressage
  - Normalisation d'URL (doublons)
  - Site ou page ?
- Politesse

Crawler - Difficultés

- Adressage
  - Normalisation d'`URL` (doublons)
  - Site ou page ?
- Politesse
  - DoS (*Denial of Service*) : DNS, Serveurs HTTP

Crawler - Difficultés

- Adressage
  - Normalisation d'`URL` (doublons)
  - Site ou page ?
- Politesse
  - DoS (*Denial of Service*) : DNS, Serveurs HTTP
  - Blacklistage officiel (`robots.txt`, `sitemap.xml`, etc.)

Crawler - Difficultés

- Adressage
  - Normalisation d'`URL` (doublons)
  - Site ou page ?
- Politesse
  - DoS (*Denial of Service*) : DNS, Serveurs HTTP
  - Blacklistage officiel (`robots.txt`, `sitemap.xml`, etc.)
  - Blacklistage officieux (*cloaking*, pièges à robot)

Crawler - Difficultés

- Adressage
  - Normalisation d'`URL` (doublons)
  - Site ou page ?
- Politesse
  - DoS (*Denial of Service*) : DNS, Serveurs HTTP
  - Blacklistage officiel (`robots.txt`, `sitemap.xml`, etc.)
  - Blacklistage officieux (*cloaking*, pièges à robot)
- Autres...

Crawler - Difficultés

- Adressage
  - Normalisation d'`URL` (doublons)
  - Site ou page ?
- Politesse
  - DoS (*Denial of Service*) : DNS, Serveurs HTTP
  - Blacklistage officiel (`robots.txt`, `sitemap.xml`, etc.)
  - Blacklistage officieux (*cloaking*, pièges à robot)
- Autres...
  - *Deep web*

Crawler - Difficultés

- Adressage
  - Normalisation d'`URL` (doublons)
  - Site ou page ?
- Politesse
  - DoS (*Denial of Service*) : DNS, Serveurs HTTP
  - Blacklistage officiel (`robots.txt`, `sitemap.xml`, etc.)
  - Blacklistage officieux (*cloaking*, pièges à robot)
- Autres...
  - *Deep web*
  - Web privé et contextualisé

Crawler

Astuces

Crawler

Astuces

- Heuristiques, tolérance

Crawler

Astuces

- Heuristiques, tolérance
- Utiliser les `headers HTTP`

Crawler

Astuces

- Heuristiques, tolérance
- Utiliser les `headers HTTP`
- User-agent

Crawler

Astuces

- Heuristiques, tolérance
- Utiliser les `headers HTTP`
- User-agent
- `random et sleep`

Crawler

Astuces

- Heuristiques, tolérance
- Utiliser les `headers HTTP`
- User-agent
- `random` et `sleep`
- Multi-agent plutôt que multi-thread

Crawler

Astuces

- Heuristiques, tolérance
- Utiliser les `headers HTTP`
- User-agent
- `random` et `sleep`
- Multi-agent plutôt que multi-thread

Principes du *Focused crawler*

Crawler

Astuces

- Heuristiques, tolérance
- Utiliser les `headers HTTP`
- User-agent
- `random` et `sleep`
- Multi-agent plutôt que multi-thread

Principes du *Focused crawler*

- Ne télécharger que les pages pertinentes

Crawler

Astuces

- Heuristiques, tolérance
- Utiliser les `headers HTTP`
- User-agent
- `random` et `sleep`
- Multi-agent plutôt que multi-thread

Principes du *Focused crawler*

- Ne télécharger que les pages pertinentes
- Indicateurs topologiques

Crawler

Astuces

- Heuristiques, tolérance
- Utiliser les `headers HTTP`
- User-agent
- `random` et `sleep`
- Multi-agent plutôt que multi-thread

Principes du *Focused crawler*

- Ne télécharger que les pages pertinentes
- Indicateurs topologiques
- Indicateurs sémantiques

Le web évolue

Le web évolue

- Le web statique devient marginal

Le web évolue

- Le web statique devient marginal
- Web dynamique (syndication : blogs, médias) : du flux RSS à PubSubHubbub

Le web évolue

- Le web statique devient marginal
- Web dynamique (syndication : blogs, médias) : du flux RSS à PubSubHubbub
- Web applicatif (réseaux sociaux, sites de contenus, micropublications)

Le web évolue

- Le web statique devient marginal
- Web dynamique (syndication : blogs, médias) : du flux RSS à PubSubHubbub
- Web applicatif (réseaux sociaux, sites de contenus, micropublications)
- Apparition de nouveaux protocoles : XMPP

Aggrégation

Principe

Aggrégation

Principe

Syndication ou comment *renverser* l'accès aux données

Aggrégation

Principe

Syndication ou comment *renverser* l'accès aux données

Avantages

Aggrégation

Principe

Syndication ou comment *renverser* l'accès aux données

Avantages

- L'information est structurée

Aggrégation

Principe

Syndication ou comment *renverser* l'accès aux données

Avantages

- L'information est structurée
- Ne capter que les nouveaux contenus

Aggrégation

Principe

Syndication ou comment *renverser* l'accès aux données

Avantages

- L'information est structurée
- Ne capter que les nouveaux contenus

Problèmes

Aggrégation

Principe

Syndication ou comment *renverser* l'accès aux données

Avantages

- L'information est structurée
- Ne capter que les nouveaux contenus

Problèmes

- Atom, RSS, encore mille versions

Aggrégation

Principe

Syndication ou comment *renverser* l'accès aux données

Avantages

- L'information est structurée
- Ne capter que les nouveaux contenus

Problèmes

- Atom, RSS, encore mille versions
- Flux complet / partiel / vide avec ou sans date, permaliens, HTML

Aggrégation

Principe

Syndication ou comment *renverser* l'accès aux données

Avantages

- L'information est structurée
- Ne capter que les nouveaux contenus

Problèmes

- Atom, RSS, encore mille versions
- Flux complet / partiel / vide avec ou sans date, permaliens, HTML
- Limitations par le propriétaire

Aggrégation

Principe

Syndication ou comment *renverser* l'accès aux données

Avantages

- L'information est structurée
- Ne capter que les nouveaux contenus

Problèmes

- Atom, RSS, encore mille versions
- Flux complet / partiel / vide avec ou sans date, permaliens, HTML
- Limitations par le propriétaire

Solution ?

Aggrégation

Principe

Syndication ou comment *renverser* l'accès aux données

Avantages

- L'information est structurée
- Ne capter que les nouveaux contenus

Problèmes

- Atom, RSS, encore mille versions
- Flux complet / partiel / vide avec ou sans date, permaliens, HTML
- Limitations par le propriétaire

Solution ?

Le paradigme publish/subscribe et l'homogénéisation

linkfluence

Les API

Principe

Les API

Principe

Utiliser les API de certains sites pour collecter la donnée

Les API

Principe

Utiliser les API de certains sites pour collecter la donnée

Avantages

Les API

Principe

Utiliser les API de certains sites pour collecter la donnée

Avantages

- L'information est *vraiment* structurée

Les API

Principe

Utiliser les API de certains sites pour collecter la donnée

Avantages

- L'information est *vraiment* structurée
- Mode de captation recommandé

Les API

Principe

Utiliser les API de certains sites pour collecter la donnée

Avantages

- L'information est *vraiment* structurée
- Mode de captation recommandé

Problèmes

Les API

Principe

Utiliser les API de certains sites pour collecter la donnée

Avantages

- L'information est *vraiment* structurée
- Mode de captation recommandé

Problèmes

- Limitations

Les API

Principe

Utiliser les API de certains sites pour collecter la donnée

Avantages

- L'information est *vraiment* structurée
- Mode de captation recommandé

Problèmes

- Limitations
- Autant de clients que d'API

Carte des profils github

- Savoir ce que l'on veut récupérer

- Savoir ce que l'on veut récupérer
- Choisir la façon la plus structurée

- Savoir ce que l'on veut récupérer
- Choisir la façon la plus structurée
- Multiplier les approches

Wikipédia est ton ami :)

- http://en.wikipedia.org/wiki/HTML
- http://en.wikipedia.org/wiki/Web_crawler
- http://en.wikipedia.org/wiki/Focused_crawler
- http://en.wikipedia.org/wiki/Web_scraping
- http://en.wikipedia.org/wiki/URL_normalization
- http://en.wikipedia.org/wiki/Cloaking
- http://en.wikipedia.org/wiki/User_agent
- http://en.wikipedia.org/wiki/Spider_trap
- http://en.wikipedia.org/wiki/Denial-of-service_attack
- http://en.wikipedia.org/wiki/Atom_(standard)
- http://en.wikipedia.org/wiki/PubSubHubbub
- etc.

Merci !

- http://labs.linkfluence.net/
- http://github.com/cmaussan/Picrowler
- http://github.com/cmaussan/captation-ic05-p10-tex

Merci !

- http://labs.linkfluence.net/
- http://github.com/cmaussan/Picrowler
- http://github.com/cmaussan/captation-ic05-p10-tex