

Guide d'utilisation du Web Crawler

ISSUE CRAWLER

1. Prérequis

Adresse du site : <http://issuecrawler.net>

Guide d'utilisation (en anglais) : http://www.govcom.org/Issuecrawler_instructions.htm

Etude de cas (en anglais) : http://www.govcom.org/scenarios_use.htm

Prérequis pour visualiser les cartes obtenues : plugin Adobe SVG
<http://www.adobe.com/svg/viewer/install/main.html>

2. Introduction : l'IssueCrawler, pour quoi faire ?

L'IssueCrawler (le « chenillard » de controverse) est un Web Crawler (« chenillard » du Web) reposant sur le concept de réseau d'affaires ou réseau de controverses. Il va permettre à son utilisateur (oui, vous !) d'utiliser le Net afin de localiser un réseau d'acteurs prenant partie dans une même controverse, et ce à travers les informations qu'ils vont partager.

Ces informations sont de nos jours de plus en plus pistables, grâce notamment à la numérisation et à la mise en ligne de tout un tas de documents, des textes bruts au format HTML, aux fichiers stockés dans des bases de données. Grâce aux hyper-liens, qui permettent de faire une sorte de « pont » entre elles, et entre les différents sites internet qui les hébergent, vous allez donc pouvoir construire un réseau des différentes éléments qui prennent part à une controverse : institutions, associations, lobbies, médias... etc, mais aussi des articles et documents divers.

Le format d'entrée va donc prendre la forme de différentes URL de la forme <http://www.siteinternet.fr/liens.html> que vous aurez choisies au préalable selon leur pertinence, après de longues et harassantes heures passées à faire des recherches sur Internet ! Elles formeront vos points de départ.

Le format de sortie va quant à lui prendre la forme d'une carte formée de clusters reliées entre-elles, chaque cluster représentant un acteur de la controverse. Ces clusters ou points auront bien sûr comme point de départ les URL rentrées précédemment, mais constitueront en bout de chaîne des points d'arrivée (pas de panique, vous comprendrez mieux tout à l'heure ;o).

C'est pour cela qu'on parle de cartographie d'un réseau d'acteurs, ou cartographie d'une controverse.

3. Le World Wide Web (WWW)

Avant tout, faisons un (tout) petit point sur la technologie web, pour comprendre le cadre

dans lequel s'insère un crawler.

Lorsque vous lisez une page d'informations sur votre navigateur (Mozilla de préférence ;o), celle-ci ne se trouve pas sur votre ordinateur, mais sur un serveur relié (connecté) à Internet par de longs et gros câbles très moches ! Donc vous « appelez » cette page, qui est ensuite traduite par votre navigateur. Différents langages permettent à celui-ci de mettre en forme les informations : HTML bien sûr mais aussi XML, XHTML, CSS ...etc.

Lorsque vous demandez à un moteur de recherche d'interroger le web, vous lui envoyez simplement une demande (requête) à laquelle il répond en interrogeant ses bases de données dans lesquelles il a au préalable stockées des informations. Ces informations sont obtenues de la manière suivante.

Les moteurs de recherche actuels, type Google ou Yahoo, sont basés sur des crawlers, c'est à dire que se sont des robots qui partent à la chasse aux liens, mais aussi qui récoltent certaines informations présentes sur les sites et qui leur sont directement destinées (pour les puristes, les balises meta description et keywords). Ces informations sont ensuite indexées dans de gigantesques base de données pour permettre une interrogation plus rapide lorsque vous saisissez votre requête.

Les crawlers sont de deux types. Tout d'abord les *smart crawlers*, qui respectent les requêtes des webmasters, c'est à dire par exemple ne pas indexer certains textes ou certaines photos. Puis viennent les *dump crawlers*, qui eux sont sans foi ni loi et ne respectent rien !

IssueCrawler utilise les deux !

Lorsque vous effectuez une requête sur IssueCrawler, vous n'interrogez donc pas une base de données prémâchées, toute belle toute propre mais bel et bien l'Internet en temps réel.

4. Le web comme terrain de jeu de la controverse

Une controverse, déjà ancienne, porte sur la véracité des informations présentes sur Internet. C'est l'éternel combat entre d'un côté la rumeur invérifiable et de l'autre l'information issue des médias dits « sérieux ».

Ce combat n'est pas le votre ! Vous devez considérer le web comme un terrain d'une richesse informationnelle très large qui va vous permettre de pister votre controverse. La distinction importante ne se fait pas sur le thème « mon info est-elle bonne ou mauvaise » mais sur le thème « mon info s'insère-t-elle dans la controverse que j'étudie ou pas ».

Savoir si l'information est bonne ou mauvaise n'est *que* (;o) le sujet plus global de votre travail sur la description d'une controverse que vous réalisez dans ce cours alors n'allez pas trop vite ni trop loin, gardez en sous le pied !

5. Où trouver la controverse sur le Web ?

Cette question revient à localiser les acteurs les plus importants de votre controverse. vous les trouverez la plupart du temps sur les sites suivants :

- les instituts scientifiques et techniques
- les pages perso des scientifiques et des ingénieurs
- les versions web des revues scientifiques et techniques
- les revues scientifiques et techniques paraissant uniquement sur internet (dites *web-only*)
- les magazines de vulgarisation sur le web
- les articles *pré-print* en ligne (par exemple : google scholar)
- les bases de données scientifiques et techniques (par exemple : web of science, medline,

google scholar)

- les sites des sociétés scientifiques (par exemple : 4S, EASST)
- les portails de l'information scientifique et technique (par exemple : INIST)

Vous y arriverez en ayant à l'esprit que pour une controverse, le web peut servir d'outil de diffusion de la controverse, mais aussi peut être une plateforme de la controverse elle-même, de l'engagement des institutions ou des acteurs sociaux dans un sens plus large.

Bien sûr, plus les sujets seront techniques, plus ils seront accessibles car circonscrits au sein d'une controverse « interne ».

Enfin, voici quelques tactiques de recherche du web:

- suivre les hyper-liens, des médias jusqu'aux archives *pré-print*, des pages perso jusqu'aux organismes financeurs
- dans Google et consors, usez et abusez de la recherche avancée et surtout de la recherche de pages similaires (pages liées et pages similaires)
- n'oubliez pas que vous avez accès depuis la bibliothèque (serveur SISTEM) à un nombre considérable de ressources en ligne
- soyez très strict sur la sélection de vos liens : il vaut mieux deux bons liens que dix mauvais liens

6. IssueCrawler : Etape n°1

Comme nous l'avons vu précédemment (bah oui, il fallait lire le début !), il va vous falloir saisir des URL, c'est à dire des adresses, qui vont pointer vers des pages web. Cela se passe dans le cadre « Harvester » (moissonner en français) de l'onglet « Issue Crawler » :

The screenshot shows the IssueCrawler web interface. At the top, there are navigation links: "Instructions of use | Scenarios of use | FAQ" and a "Log Out" link. The main header displays "issuecrawler". Below the header is a menu with four tabs: "the Lobby", "Issue Crawler" (highlighted with a yellow circle), "Network Manager", and "Archive". The main content area is divided into two columns. The left column contains the "Harvester" section (highlighted with a yellow circle). It includes a text input field with the instruction "Type or paste text and URLs into the Harvester" and a note: "The text will be stripped to create starting points for the Issue Crawler". Two URLs are entered in the field: "http://www.site1.fr/page_de_liens1.html" and "http://www.site2.fr/page_de_liens2.html" (both highlighted with yellow circles). Below the input field is a "Harvest" button (highlighted with a yellow circle). To the right of the input field is a "Next step >>" link with the text "Fine tune and Launch Crawl". The right column contains a "Current Crawls" section with an "RSS" link. The interface also shows a date "Wednesday, January 04, 2006" and a user identifier "@683.46".

Puis cliquer sur « Harvest ».

Quelques astuces :

- si vous avez la chance de tomber sur LE site qui contient une page entière de liens hyper super pertinents à moissonner, aller dans le menu « Edition » de votre navigateur préféré, sélectionner « voir la source », sélectionner toute la page et coller dans le cadre « Harvester ». quand vous cliquerez sur le bouton « Harvest », le logiciel sélectionnera et nettoiera automatiquement les URLs.
- ce qui est au-dessus n'est néanmoins pas très raisonnable car vous allez vous retrouver avec une carte très dense et peut être illisible ! Préférez commencer avec 2 ou 3 URL, c'est largement suffisant pour tâter le terrain

7. IssueCrawler : Etape n°2

Voici venu le temps des paramétrages du crawler. La consigne principale est la suivante : pour la majorité des crawls, les options par défaut seront les meilleures. Simplissime non ? (... ah ces concepteurs :o)

Quelques explications tout de même :

- Privilege starting points (points de départ privilégiés) : ce critère vaut lorsque vous êtes sûr du point de départ, c'est à dire lorsque vous êtes sûr de vos acteurs
- Perform co-link analysis by site or by page (analyse des co-liens par page ou par site) : vous retourne un réseau de site (par leur home page) ou un réseau de page
- Set iteration to 1,2 or 3 :
 - choisissez 1 pour localiser le réseau social d'une organisation à partir d'url d'autres organisations relatives,
 - 2 pour localiser un réseau d'organisations à partir de pages relatives à une controverse,
 - 3 pour un réseau plus large d'une sphère politique.
- Set crawl depth to 1,2 or 3 (profondeur du crawl) : les pages visitées à partir de vos points de départ sont considérées comme la « profondeur 0 »
 - choisissez 1 si vous voulez visiter les pages visitées par le premier crawl
 - choisissez 2 si vous voulez visiter les pages visitées par le second crawl
 - choisissez 3 si vous voulez visiter les pages visitées par le troisième crawl
- Paramètres avancées : (je vous conseille de les laisser tels quels car ils suffisent pour la majorité des crawls)
 - Crawled URL ceiling (per host) : nombre maximum d'URL crawlée sur chaque hôte, c'est à dire sur chaque site visité
 - Crawled URL ceiling (overall) : quantité totale d'URL crawlée
 - Co-link ceiling by page (pages per host per iteration) : quantité maximum de pages liées retournées par itération
 - Co-link ceiling by site (hosts per iteration) : quantité maximum de sites liés retournés par itération
- Exclude from Network : c'est la liste d'exclusion du crawl, c'est à dire les sites ou pages que vous ne voulez pas voir apparaître sur votre carte finale. En général, vous complèterez la liste une fois votre premier crawl effectué. Sinon, elle est constituée principalement des moteurs de recherches, des pages de téléchargement de soft, des gros sites type Microsoft ...etc.

http://www.site1.fr/page_de_liens1.html
http://www.site2.fr/page_de_liens2.html

View Edit Update Remove Save Results

Privilege starting points Use privilege starting points and one iteration to find an immediate social network.	<input checked="" type="radio"/> Off (default) <input type="radio"/> On
Perform co-link analysis by Use co-link analysis by page for more specific results. The nodes are likely to be deep pages.	<input type="radio"/> Site <input checked="" type="radio"/> Page (default)
Set iterations Use at least two iterations for locating an issue network.	<input type="radio"/> 1 <input checked="" type="radio"/> 2 (default) <input type="radio"/> 3
Set crawl depth Use a deeper crawl depth for homepages.	<input type="radio"/> 1 <input checked="" type="radio"/> 2 (default) <input type="radio"/> 3

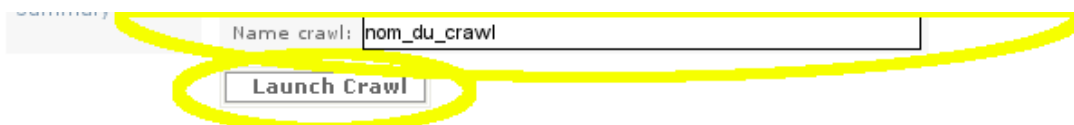
Ceilings (advanced):

Crawled URL ceiling (per host) The maximum number of URLs crawled on each host.	<input type="text" value="500"/>
Crawled URL ceiling (overall) The maximum number of URLs crawled, (max 60000)	<input type="text" value="40000"/>
Co-link ceiling by page (pages per host per iteration) The maximum quantity of co-linked pages returned per iteration, (max 1000)	<input type="text" value="100"/>
Co-link ceiling by site (hosts per iteration) The maximum quantity of co-linked sites returned per iteration, (max 1000)	<input type="text" value="100"/>

Exclude from network:

download.cnet.com ;
download.com ;
download.net ;
netscape.com ;

Donnez un nom à votre crawl puis cliquez sur « Launch crawl »



A screenshot of a web form. The form has a label 'Name crawl:' followed by a text input field containing the text 'nom_du_crawl'. Below the input field is a button labeled 'Launch Crawl'. A yellow oval highlights the 'Launch Crawl' button, and a larger yellow oval highlights the entire form area.

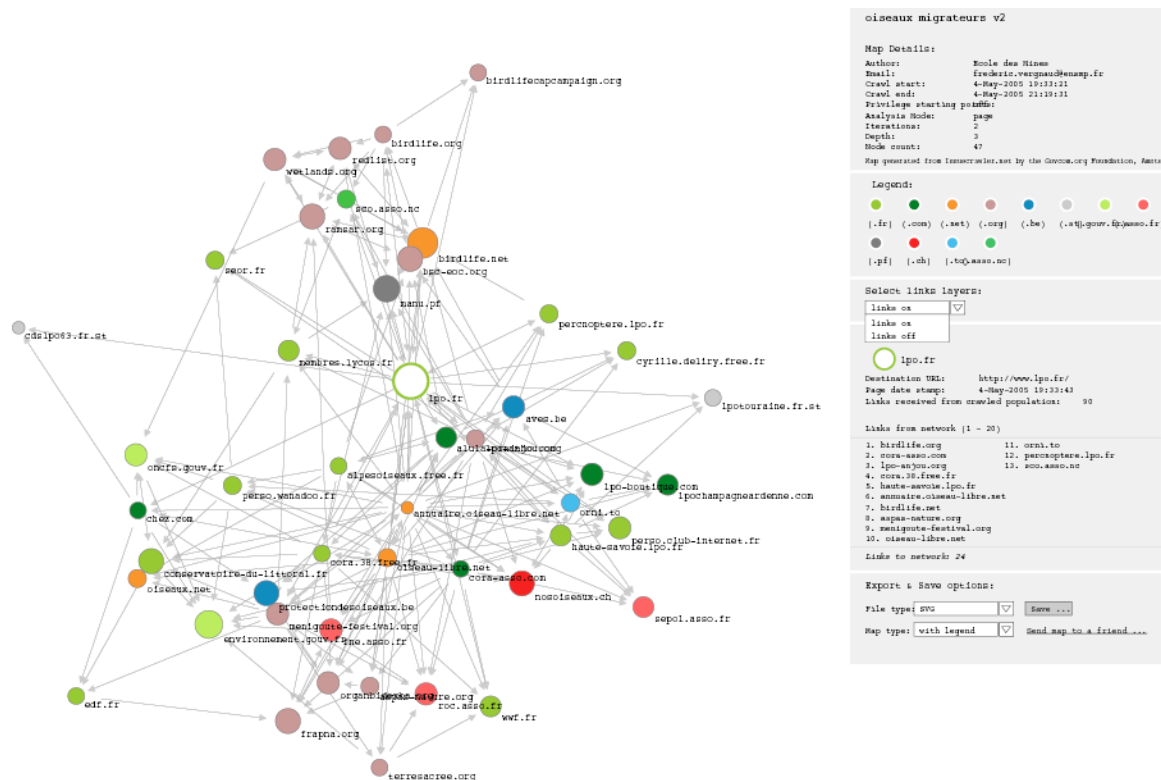
Vous n'avez plus qu'à attendre !

7. IssueCrawler : Etape n°3

Cette étape va vous permettre de visualiser votre crawl, c'est à dire le réseau que vous avez obtenu sous forme de carte.

Cliquez simplement sur le nom de votre crawl puis sur « View depiction » et attendez le chargement de la page (le plugin *ADOBE SVG* doit être installé sur votre bécane pour visualiser la carte).

Voici à quoi cela ressemble :



Oh la belle carte !

La carte

- Tous les noeuds / noms de sites/pages sont cliquables et ouvrent une fenêtre popup où s'affiche la page cliquée
- vous pouvez zoomer sur une partie du réseau en faisant CTRL droit de la souris

Le côté droit

- vous pouvez afficher ou faire disparaître certains domaines en cliquant dessus
- de même pour les liens, par le petit menu en dessous « links on », « links off »
- vous pouvez enregistrer votre carte au format SVG, avec ou sans les légendes. Elle sera directement intégrable dans votre site de controverse et gardera toutes les propriétés

susmentionnées.