UMONS
Université de Mons

[Session « Démo »]
« *Spygraph : un robot d'exploration léger dédié à l'analyse de graphes d'hyperliens* »

Dr Ir Robert Viseur
Chargé de cours

Congrès Inforisid 2022

Université de Bourgogne – 01 juin 2022

Interreg France-Wallonie-Vlaanderen FabricAr3v

Avec le soutien du Fonds Européen de développement régional –
Met steun van het Europees Fonds voor Regionale Ontwikkeling.

ISO 9001 VINÇOTTE

POLYTECH MONS

Motivations

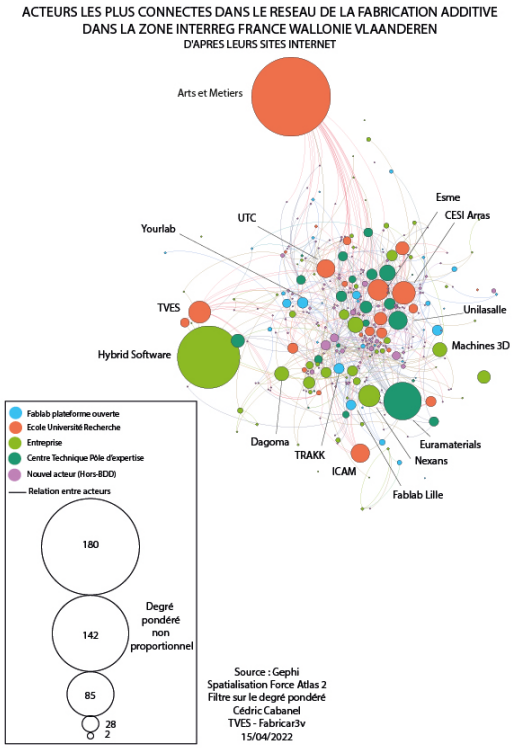
Disposer d'un robot d'exploration léger et configurable permettant :

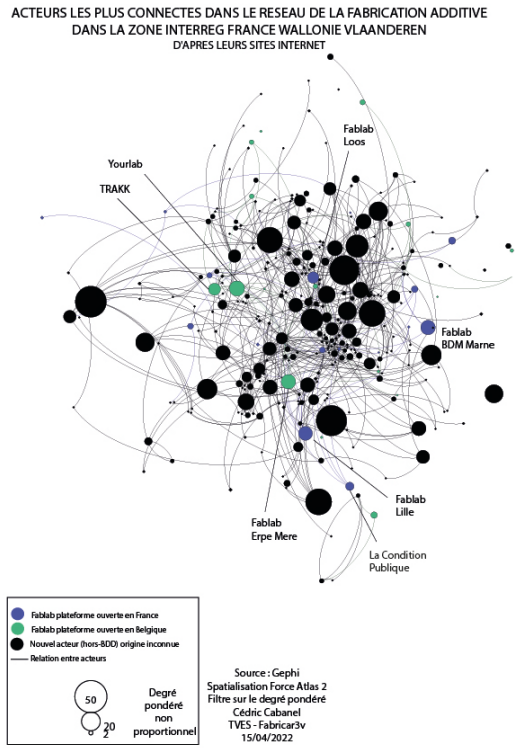
- l'exploration d'un écosystème d'affaires local au travers des pages web,
  (p. ex. découverte de nouveaux opérateurs)
- la production de données pour l'analyse de graphes d'hyperliens.
  (p. ex. exportation vers Gephi ; cf. Hansen et al., 2010)

Contraintes imposées (positionnement) :

- centrage sur l'exploration des sites,
- respect de la philosophie KISS (*Keep It Simple, Stupid*),
- compatibilité avec les outils d'analyse de graphes (p. ex. Gephi),
- complémentarité aux logiciels pour grands graphes (p. ex. Hyphe).

Application au projet « FabricAr3v »





Crédits : Cédric Cabanel <cedric.cabanel@univ-lille.fr>.

Utilisation du programme

Le programme se lance en ligne de commande :

```
#!/bin/bash
python3 run.py init <monprojet>
python3 run.py crawl inside <monprojet>
python3 run.py export csv <monprojet>
python3 run.py export graph <monprojet>
```

Le programme a été testé sous Ubuntu Linux.

Le logiciel est programmé en Python.

Les dépendances incluent Beautiful Soup.

La base de données est SQLite.

Configuration du programme

La configuration est gérée, d'une part, via les paramètres de la ligne de commande, d'autre part, via un fichier de configuration (`run.cfg`).

Paramètres :

`init`	Création de la base de données et initialisation de la table « urls_init » avec la liste d'URLs d'amorçage.	
`crawl`	Exécution de l'exploration sur base de la liste d'URLs.	
	`inside`	Exploration des URLs associées aux domaines fournis dans la liste d'amorçage uniquement.
	`outside`	Exploration de toutes les URLs trouvées (mode divergent).
`export`	Exportation des données issues de la base de données une fois l'exploration terminée.	
	`csv`	Exportation des liens entre pages au format CSV.
	`graph`	Exportation des liens entre pages au format DOT.

Configuration :

`limit`	Nombre maximal d'URIs explorées.
`depth`	Profondeur de l'exploration.
`stealth`	Fonctionnement « discret » ou non.
`timeout`	Délai de timeout à l'ouverture d'une page
`sqldot`	Filtrage pour l'exportation du graphe d'hyperliens (SQL).
`sqlcsv`	Filtrage pour l'exportation du tableau de données (SQL).

Structure de la base de données

La base de données comporte trois tables :

urls_init	urls	links
url fulldomain domainname	url fulldomain domainname iscrawled iserror isignored	hyperlink_from hyperlink_to fulldomain_from fulldomain_to domainname_from domainname_to

Avec :

`url_inits`	URLs permettant d'amorcer l'exploration.
`urls`	Liste des URIs découvertes.
`links`	Liste des liens découverts.

Publication

Le code source a été publié en ligne (Spygraph) sous licence GPL v3.

Les contributions sont les bienvenues :

- les retours d'expérience (positifs ou négatifs),
- les comparaisons à d'autres logiciels,
- les rapports de *bugs*,
- les suggestions de modifications,
- les codages de nouvelles fonctionnalités…

Contacter Robert VISEUR <robert.viseur@umons.ac.be>.

En cas d'utilisation, la référence à l'article de présentation sera appréciée : « *Robert Viseur (2022), Spygraph : un robot d'exploration léger dédié à l'analyse de graphes d'hyperliens. Inforsid, Dijon (France)* ».

Perspectives

Perspectives à court et moyen terme :

- Générer des rapports d'exploration.
- Rajouter un mode `strict` au *crawl*.
- Améliorer la fonctionnalité de filtrage (exportation).

Perspectives à long terme :

- Extraire automatiquement, pour chaque page, un ensemble d'expressions, puis :
  - Permettre un filtrage par « mots »-clefs.
  - Produire des nuages de « mots »-clefs.
- Compléter le robot d'exploration par un indexeur (p. ex. PyLucene).
- Accélérer l'exploration par « *focused crawling* » (cf. Micarelli et al., 2007).

Cf. Viseur (2012, 2014a, 2014b) pour des exemples.

Démonstration

```
Fichier  Édition  Affichage  Rechercher  Terminal  Aide
  Starting 7876 : https://astertechnics.be/inschrijving (1, 1).
  Starting 7877 : https://astertechnics.be (1, 1).
  Starting 7878 : https://astertechnics.be/het-beestig-bosspel (1, 1).
  Starting 7879 : https://astertechnics.be/activiteiten (1, 1).
  Starting 7880 : https://astertechnics.be/workshops-op-maat (1, 1).
  Starting 7881 : https://astertechnics.be/activiteiten%20-%20algemene%20info (1, 1).
  Starting 7882 : https://www.arteveldehogeschool.be/projecten/lets-stem-together (1, 1).
  Starting 7883 : https://astertechnics.be/repair-cafe-3 (1, 1).
  Starting 7884 : https://astertechnics.be/2021/10/20/jonge-ontdekkers-poppen-maken-met-herfstmaterialen-20-10-2021/ (1, 1).
  Starting 7885 : https://astertechnics.be/category/foto/ (1, 1).
  Starting 7886 : https://astertechnics.be/2021/08/13/fotografiekamp-09-13-08-2021/ (1, 1).
  Starting 7887 : https://astertechnics.be/2021/07/16/kookkamp-12-07-t-e-m-16-07-2021-2/ (1, 1).
  Starting 7888 : https://astertechnics.be/voorbeeld-pagina/ (1, 1).
  Starting 7889 : https://astertechnics.be/historiek (1, 1).
  Starting 7890 : https://astertechnics.be/nieuwesite/robotje (1, 1).
  Starting 7891 : https://astertechnics.be/nieuwesite/infrastructuur (1, 1).
  Starting 7892 : https://astertechnics.be/nieuwesite/missie (1, 1).
  Starting 7893 : https://astertechnics.be/nieuwesite/www.astertechnics.be/inzamelpunt (1, 1).
  Starting 7894 : https://astertechnics.be/nieuwesite/onze-partners (1, 1).
  Starting 7895 : https://www.fablaberpemere.be (1, 1).
  Starting 7896 : https://www.fablaberpemere.be/video1 (1, 1).
  Starting 7897 : https://www.fablaberpemere.be/open-fablab (1, 1).
  Starting 7898 : https://www.fablaberpemere.be/workshops (1, 1).
  Starting 7899 : https://www.fablaberpemere.be/kopie-van-open-fablab (1, 1).
  Starting 7900 : https://www.fablaberpemere.be/fablab-kampen (1, 1).
  Starting 7901 : https://www.fablaberpemere.be/steam-clubs (1, 1).
  Starting 7902 : https://www.fablaberpemere.be/video-jongeren (1, 1).
  Starting 7903 : https://www.fablaberpemere.be/machines (1, 1).
  Starting 7904 : https://www.fablaberpemere.be/blog-1 (1, 1).
  Starting 7905 : https://www.fablaberpemere.be/agenda (1, 1).
  Starting 7906 : https://www.fablaberpemere.be/info-vzw (1, 1).
  Starting 7907 : https://www.fablaberpemere.be/contact (1, 1).
  Starting 7908 : https://www.fablaberpemere.be/privacy-policy (1, 1).
  Depth level 2 is reached.
  The crawl was completed (01h 34min 37s).
End.
Exporting the files...
  The file with URLs exists.
  The file of database exists.
  The CSV file was exported (8.56s).
End.
Exporting the files...
  The file with URLs exists.
  The file of database exists.
  The graph file was exported (0.23s).
End.
rv@lt-rv-2015:/media/rv/SSD_DATA_2021/transitional/dev/experimentation/spygraph$
```

Partenaires du projet

Interreg France-Wallonie-Vlaanderen
FabricAr3v

Cenaero

centralelille

cnrs

CRITT Matériaux Innovation

ISOLECTRA martin

materalia

Machines-3D

sirris
driving industry by technology

UMONS

Université de Lille

Projet soutenu par

Interreg France-Wallonie-Vlaanderen

Recherche et innovation

www.interreg-fwvl.eu
@InterregFWVL

Wallonie

Avec le soutien du Fonds européen de développement régional

UMONS
Université de Mons

Université de Mons
Faculté Warocqué d'économie
et de gestion - Service TIC
Place Warocqué, 17
B-7000 Mons

Tél. : +32.65.373.201

www.umons.ac.be
info.warocque@umons.ac.be

Plus d'information...

Dr Ir Robert VISEUR
Chargé de cours

Tél. : +32.65.374.054
robert.viseur@umons.ac.be