Graphe du Web

Pour ce TP nous allons nous intéresser au graphe du Web. Nous utiliserons le language java (version 1.4.2)[1]*. Les objectifs du TP sont les suivants :*
- *Être capable de faire l'exploration d'une partie du Web ;*
- *Sauvegarder le graphe du web que vous aurez exploré dans un fichier xml ;*
- *Évaluer quelques propriétés caratéristiques du graphe.*

Vous trouverez de l'aide sur la page web `http://graal.ens-lyon.fr/~rbolze/art.html`

1 Exploration de la toile du Web

Définition 1. *Le graphe du web peut être modélisé par un graphe orienté dont les sommets sont des pages web et un arc représente un lien hypertextuel qui pointe d'une page vers une autre.*

La première étape consiste à traiter une page Web afin d'extraire les liens hypertexte pour ensuite explorer la "toile", à la manière des robots d'indexation (ou araignées ; en anglais web crawler ou web spider).

Question 1.1. *A l'aide de l'exemple* TitleExtractor.java *[**?**], écrivez une classe* LinkExtractor.java *qui permet d'obtenir la liste des liens d'une page web.*

Vous pouvez utiliser comme page de test de vos programmes l'url suivante :
`http://graal.ens-lyon.fr/~rbolze/linkExtractor/index.html`

Question 1.2. *Ecrivez une classe* Spider.java *qui fait l'exploration du graphe du Web. Posez vous les questions suivantes :*

- *Quels sont vos critères d'arrêt de l'exploration ?*
- *Quelles sont les propriétés a priori de ce graphe ?*
- *Comment allez-vous coder ce graphe ?*

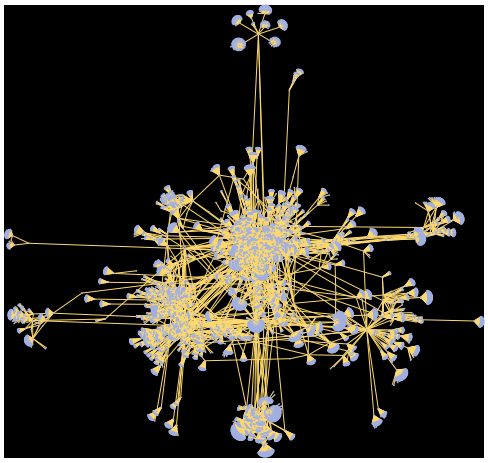

FIG. 1 – graphe du web généré à partir de www.cnrs.fr

[1] version disponible sur les machines de la salle, rien ne vous empêche d'utiliser une version plus récente

2 Sauvegarder la toile

Maintenant que vous savez extraire le graphe du web. Je vous propose de sauvegarder votre exploration dans un fichier au format graphML [**?**].
Vous trouverez un exemple *WebGraphML-example.graphml* sur la page
`http://graal.ens-lyon.fr/~rbolze/art.html`.

Question 2.1. *Ecrivez une classe* GraphMLConverter.java *permettant la sauvegarde de votre exploration dans un fichier au format graphML [**?**].*

Servez vous des logiciels yED [**?**] et GUESS [**?**] pour visualiser vos graphes.

Question 2.2. *Réutiliser la classe* GraphMLScanner.java *afin de lire les fichiers au format graphML que vous avez sauvegarder lors de vos explorations du Web.*

3 Propriètés du graphe du Web

Vous avez vu dans le cours :

- **La distance moyenne** ; (de l'ordre de 19)
- **Le coefficient de clusterisation** ; (coefficient élevé)
- **La distribution des degrés** ; (loi de puissance $d^{-\alpha}$ avec $\alpha > 1$)

Question 3.1. *Évaluer ces trois propriétés pour les graphes que vous avez explorés. Que remarquez-vous ?*

Définition 2. *La distance moyenne d'un graphe est la moyenne des distances entre toutes les paires de sommets.*

$$\bar{dist} = \sum_{u \neq v} \frac{dist(u,v)}{C_n^2}$$

Définition 3. *Il existe deux définitions pour le coefficient de clusterisation*

– *une définition globale :*

$$C_1 = \frac{\sum_{u \in V} \text{nombre de triangles dont } u \text{ est un sommet}}{\sum_{u \in V} \text{nombre de triplets dont } u \text{ est un sommet}}$$

– *une définition basée sur la moyenne de coefficients locaux :*

$$C_2 = \frac{1}{n} \sum_{u \in V} \frac{\text{nombre de triangles dont } u \text{ est un sommet}}{\text{nombre de triplets dont } u \text{ est un sommet}}$$

Sources et références