Le filtrage se fera du coté serveur et pourra etre couplé avec <http://www.clamav.net/> ?

Le client devra s’authentifier et s’inscrire sur un site ?

Question : chaque crawler définit dans l’interface permet de crawler une adresse URL.

L’application coté client ne devra pas être trop lourde et trop complexe à configurer. L’utilisateur cliquera sur start et le crawl se lancera à partir de sa machine sans qu’il n’est rien à configurer. Il pourra cepandant si il le veut configurer le logiciel. Les options de config seront le nombre de threads sur la machine

On décide de faire du crawl par catégorie.

Scenarios

**Côté Client :**

**Scenario 1 : Lancement du crawl**

Bob souhaite démarrer un crawl, il s’authentifie avec son login. Arrivé sur l’interface graphique il décide de lancer un crawl, il clique alors sur le bouton Start. Le logiciel se lance en tâche de fond et les différents liens crawlés sont visible pour Bob.

**Scenario 2 : Paramétrage du crawl**

Bob souhaitant paramétrer son crawler clique sur le bouton options, alors apparait un certain nombre de paramètres. Il souhaite mettre 20 threads de connexion fonctionnant simultanément, puis il met la quantité d’information téléchargé par lien à 20Mo, et (enfin règle la vitesse de download maximum à 50Ko/s et d’upload à 20 Ko/s).

**Scenario 3 : Options avancés du paramétrage du crawl**

Bob décide d’aller plus loin dans la configuration du crawl il clique alors sur paramètre avancé, il se rend ensuite sur l’onglet proxy ou il rentre les informations nécessaire au bon fonctionnement de la connexion client serveur. Il se rend ensuite sur l’onglet des identifiants et vérifie si ils sont corrects pour démarrer le crawl.

**Scenario 4 : Statistiques**

Bob décide de voir l’état du crawl en temps réel, il clique alors sur statistiques, il voit alors apparaitre le temps de connexion sur le serveur, le nombre de crawler actif, la vitesse de download et d’upload, le nombre de liens crawlé, les erreurs.

**Scenario 5 : Détails du crawl**

Bob décide de regarder tous les crawlers en action, il clique alors sur details. Une liste des crawlers apparait avec leurs status, leurs identifiants, l’avancé du download en pourcentage, le nombre de Mo downloadé.

**Scenario 6 : Visualisation du graphe**

Afin de voir la carte des différents sites crawlé, Bob décide de cliquer sur sur voir graphe. Il visualise alors la carte des différents sites crawlés par lui sous forme de noeud. Il choisit ensuite une catégorie dans la liste fournit par le serveur et il clique sur générer graphe afin de voir le graphe correspondant à celle-ci.

**Scenario 7 : Navigation dans le graphe (3 couches)**

Bob souhaite accèder à un nœud du graphe ou l’agrégat lui parait conséquent, il clique alors sur ce nœud. Des infos relatives au nœud s’affiche telle que l’URL, la profondeur de celle-ci, le nombre de liens entrant et sortant sur ce nœud. Il décide ensuite d’aller à cette adresse en cliquant sur go dans l’interface ou s’est affiché les informations sur le nœud.

**Côté Serveur :**

**Scenario 1 : Entrée des URLs**

Alice souhaite intégrer une liste des adresses à crawler, elle clique alors sur charger URL et ouvre son fichier afin qu’il se charge dans l’interface. Alice se rend compte qu’elle à oublié une URL, elle clique alors sur ajouter URL, puis retire une URL qui ne sert finalement à rien en cliquant sur supprimer URL. Enfin elle choisit d’exporter la liste.

**Scenario 4 : Filtrage des URLs**

Alice ne désire pas voir apparaitre dans le résultat du crawl certaines adresses qui pourrait parasiter la pertinence de la recherche. Elle clique alors sur ajouter URL et entre

**Scenario : Graphe du serveur**

Comme le client -> graphe final serveur

3 niveaux :

Le graphe de site : le siteduzero -> developpez

Le graphe d’URL : lesiteduzero/java

lesiteduzero/php

1er niveau

[www.lesiteduzeo.com](http://www.lesiteduzeo.com) : www est le sous domaine du graphe de site

2eme niveau

forum.siteduzeo.com : forum est le sous domaine

news.siteduzero.com news

3eme niveau est l’url

forum.siteduzero.com/index.php?topic=1234&page=456

4eme niveau

But final

Perspective d’évolution

**Scenario 8 : Création d’un sitemap XML**

Afin de contribuer à la pertinence et la rapidité de recherche de l’information, Bob décide de créer un sitemap, il clique alors sur Sitemap et entre l’URL qu’il veut mapper. Il appuie ensuite sur envoyer pour crawler cette URL, un fichier est ensuite crée et déposé sur le serveur.

**Sécurité**

Question : L’application serveur par n’importe qui ??? Peut être si authentification pour disposer de la version serveur, afin de gérer les adresses crawlés

**Scenario 2 : URL par mots clefs**

Alice souhaite définir des mots clés qualifiant la liste des URLs qu’elle a chargé. Elle clique alors sur mots clefs et ajoute un mot cléf en lui indiquant un score de 0 à 10 sur la pertinence du mot.

**Scenario 3 : URL par catégories**

Alice désire filtrer par catégories, le but étant pour elle de classifier les adresses et de rendre plus clair son filtre. Elle entre donc des adresses qu’elle ne voudra pas voir et créer une catégorie pour celle-ci. Elle décide ensuite d’exporter cette catégorie au format XML.

**Scenario 5 : Paramétrage du serveur**

Alice souhaite paramétrer -> Comment ca marche : le serveur doit connaitre qui ? (les clients et les serveur rattachés à lui !!!)

Quelles clients décide de prendre le serveur et comment il fait son choix ? Sur le client il y aura une liste de serveur auquel il peut se connecter et cette liste sera récupéré par un serveur central dont son role sera de récupérer les adresses des serveurs et les distribué au client.

Uml :

Bouml