1 Étapes déjà réalisées et/ou en cours

Étude bibliographique et idée d'une solution
Lecture de plusieurs papiers traitant de sujets similaire. C'est pendant cette phase que j'ai eu l'idée d'utiliser un algorithme de génétique car ils existent depuis un moment et il en existe qui permettent de détecter des anomalies génétiques, ce qui se rapproche du problème du projet.

Récupération de données brutes pour la création d'historique de navigation
Pour ce projet il me faut des historiques de navigation internet et comme il n'est pas aisé d'en obtenir des vrais en assez grande quantité. J'ai donc fait un programme java qui permet de récupérer toutes les infos dont j'ai besoin sur le site *alexa.com* qui est un site classant les différents sites internet et fournissant des informations sur ces sites, en particulier les 10 premiers parents dans la navigation avec leur probabilité, c'est à dire pour un site donné les 10 site précédent dans la navigation les plus probable.

Le programme effectue plusieurs centaines de requêtes http pendant une dizaine de minutes et stockes toutes les informations pour les 500 premiers sites les plus visités en France dans un fichier xml.

Création d'un groupe d'individu test et de leurs historiques
Le deuxième programme java demande à l'utilisateur d'entrer un nombre d'individus pour le groupe test ainsi qu'une profondeur maximale et minimale pour les historiques de navigation et un nombre minimal et maximal de racines d'arbre de navigation.

Pour chaque individu du groupe, le programme choisi un nombre de racine entre le min et le max parmi les 50 premiers sites français puis il choisi un parent parmi les 10 à ajouter à l'historique en faisant un tirage au sort et en fonction des probabilités de chacun des parents.

On choisi des parents jusqu'à ce que l'on ait atteint la profondeur pour la racine en question (choisi elle aussi entre le min et le max) ou bien on s'arrête si un parent choisi n'a pas de parent connu (ie il n'est pas dans la liste des 500 premiers sites dont on connaît les parents). Dans ce cas on a une chance sur trois de choisir un retour sur google (ie on prend google comme parent). Sinon on arrête la branche pour cette racine et on passe à la racine suivante.

Tri des historiques de navigation par catégories
Pour cette étape j'ai modifié le premier programme pour qu'il récupère également les informations de catégories des 500 sites.

Pour le second programme, une fonction *sort* permet de trier les historiques une fois qu'ils sont créés en fonction de leur catégorie.

Pour l'instant j'en suis ici. Je n'ai les informations de catégorie que pour les 500 premiers site mais je compte modifier le programme afin de récupérer en ligne en direct pendant le tri les catégories des sites qui ne figure pas dans les 500 premiers.

2 Étapes à venir

Faire tourner l'algorithme de génétique
Une fois les sites triés par catégories, il faudra définir des fonctions de notation de risque pour chacune des catégories et implémenter l'algorithme de génétique qui est déjà utilisé en informatique pour détecter des intrusions sur un système. Le papier décrivant cet algorithme est fourni avec ce document.

Évaluation de la pertinence des résultats
Une fois les résultats obtenus en sortie de l'algorithme, il faudra les évaluer pour juger de la qualité de la détection et donc ensuite en cas de problèmes, soit affiner les fonctions de natation du risque, soit changer d'algorithme de détection.