Indexation et recherche de contenu utilisant MapReduce de Google

Documentation utilisateur

Elodie Corbel, Kévin M'Ghari,
Mickaël Olivier, Clarisse Renou

Encadrant : Alexandru Costan

Introduction

Notre projet intitulé Indexation et recherche de contenu utilisant MapReduce de Google est en fait un moteur de recherche. Nous utilisons Hadoop pour le système d'indexation, c'est-à-dire référencer les fichiers dans notre moteur de recherche et une applet Java pour effectuer la recherche. Cette documentation a pour but de vous expliquer comment installer et utiliser notre moteur de recherche.

1 Installation

Pour l'installation, il y a plusieurs logiciels nécessaires au fonctionnement de notre moteur de recherche.

1.1 Pré-requis

Tout d'abord, un système Unix est nécessaire pour l'exécution des scripts.

Ensuite, il faut avoir le Java Development Kit (JDK) installé sur votre ordinateur pour faire fonctionner notre logiciel.

Enfin, il vous faudra installer et configurer Hadoop pour avoir accès à l'indexation. Pour plus d'informations, aller voir la documentation officielle : `http://hadoop.apache.org/`. Si vous voulez suivre un tutoriel pour l'installation sur Ubuntu, c'est ici : `http://www.michael-noll.com/tutorials/running-hadoop-on-ubuntu-linux-single-node-cluster/`.

1.2 Configuration

Pour configurer notre moteur de recherche, plusieurs fichiers sont à modifier.

Pour le système d'indexation, il est impératif de modifier le fichier `scripthadoop.sh` se trouvant dans l'archive de notre projet dans le dossier `hadoopMR`. Les variables à modifier sont :

- `hadoopPath` chemin absolu où vous avez installé Hadoop
- `inputPath` chemin absolu du dossier parent où se trouvent vos fichiers que vous voulez indexer, les fichiers à indexer doivent obligatoirement être dans un sous-dossier de ce répertoire nommé `inputFiles`
- `dfsInputPath` endroit où vous voulez mettre les fichiers sur le système de fichiers Hadoop

- `dfsOutputPath` l'endroit où vous voulez mettre les fichiers en sortie dans le système de fichiers d'Hadoop, le dossier ne doit pas exister au préalable sinon vous aurez une erreur au lancement.

Si vous voulez que l'indexation s'effectue automatiquement toutes les heures, il faut planifier une tache cron sur Unix[1]. Sinon, exécutez seulement le script `scripthadoop.sh` dans le dossier `hadoopMR` du projet.

Pour ce qui est de l'applet, vous allez devoir modifier le code et recompiler notre applet (en gardant les mêmes noms et en la plaçant au bon endroit). Pour accéder au code source de notre applet, dans notre archive, vous le trouverez dans le dossier `SearchEngine`. Il vous faut ensuite aller dans `src/path/Paths.java`. Il faut modifier chacune des variables. Normalement,

- `outputIndexLocation` même contenu que inputPath puis rajouter `/outputFiles/output`
- `inputFilesSplitDir` même contenu que inputPath puis rajouter `/inputFilesSplit/`
- `logFilePath` chemin où vous voulez que soit placé le fichier journal du programme

Puis, compilez et placez le .jar de l'applet dans le dossier `Page web` en la nommant applet.jar. N'oubliez pas de signer l'applet (obligatoire pour lire les fichiers du disque dur et donc pour parcourir l'index).

2 Fonctionnalités

Après les étapes complexes de configuration, passons au fonctionnement du moteur de recherche en lui-même. Pour lancer le logiciel, il vous suffit, après avoir exécuté le script d'indexation (voir section Installation), d'ouvrir, avec un navigateur web, le fichier `page.html` se trouvant dans le dossier `Page web` de notre projet.

2.1 Recherche simple

Pour faire une recherche, il vous faut simplement entrer des mots dans le champ en-dessous du petit éléphant jaune (logo d'Hadoop), appuyer sur Entrée (ou Search)et le résultat vous sera donné. Le résultat se constituera des noms de fichiers dans lequel votre(vos) mot(s) se trouve(nt) et des trois lignes l'entourant dans le fichier, voir aperçu figure 1.

A noter que certains mots non pertinents sont ignorés et ne donnent pas de résultats. Voici la liste de ces mots : *et, ou, où, de, des, d, le, les, l, la, je, il, au, aux, du, un, une, a, à, ni, que, si, y, m, mon, ma, mes, me, ne, nous, on, sa, ses, se, qui, s, t, ta, tes, te, il, là, qu, sans, sur.*

2.2 Recherche avec prédicats

Vous avez la possibilité d'effectuer une recherche avec des prédicats dans notre moteur de recherche. Les prédicats `AND`, `OR` et `NOT` ont une signification.

Le prédicat `AND` permet de rendre en résultat seulement les fichiers où les 2 mots suivant le prédicat apparaissent. Syntaxe de la commande : `AND mot1 mot2`.

Le prédicat `OR` permet de réaliser un ou exclusif. Il rendra seulement les fichiers où les 2 mots suivant le prédicat n'apparaissent pas ensemble. Syntaxe de la commande : `OR mot1 mot2`.

1. voir manuel Unix

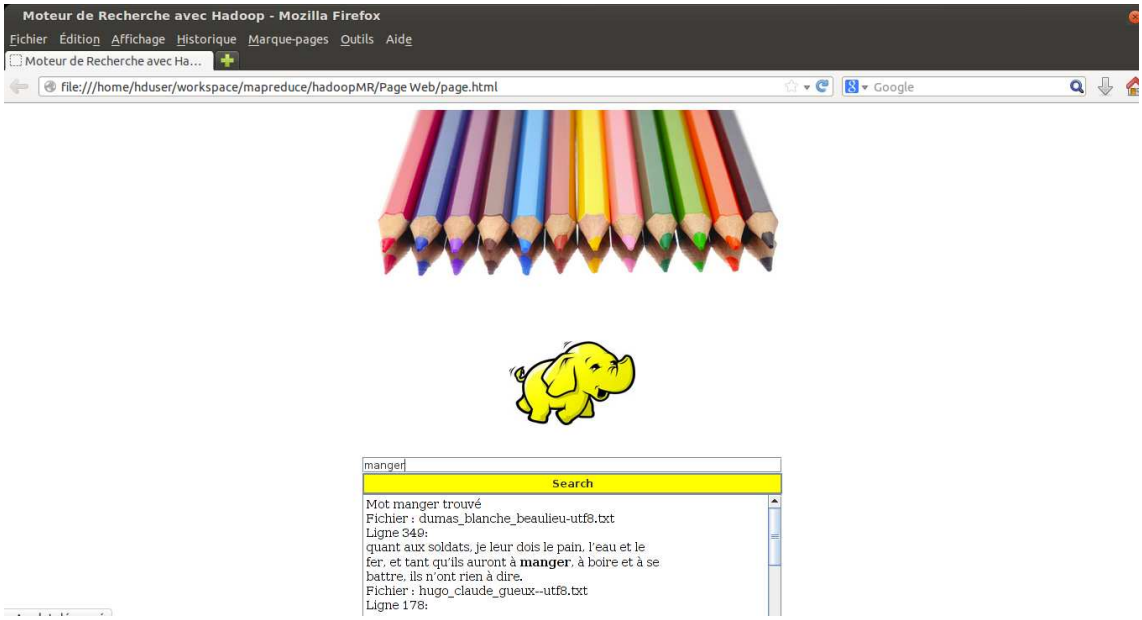

FIGURE 1 – Apercu de notre moteur de recherche

Le prédicat `NOT` permet d'exclure les fichiers contenant le mot suivant le prédicat. Syntaxe de la commande : `NOT mot`.

3 Conclusion

Notre moteur de recherche est donc assez simple. Il peut même être installé sur un serveur de type Wamp. Il suffit juste de mettre le contenu de l'archive dans le dossier du serveur. Si vous avez des questions, n'hésitez pas à nous contacter.