POLYTECH PARIS-UPMC

MAIN5

Xavier Tannier

Projet : Moteur de Recherche d'Information

1 Objectif

Le but de ce projet est de réaliser un moteur de recherche d'information.

- Entrée du système : une requête composée d'un ou de plusieurs mots-clés représentant le besoin d'information d'un utilisateur.
- Sortie du système : une liste ordonnée de pages web répondant au besoin de l'utilisateur.

Le fonctionnement attendu est donc le même qu'un moteur de recherche tel que Google, Bing ou Qwant [1].

2 Description générale

Ces moteurs fonctionnent globalement selon l'architecture suivante, illustrée à la figure 1 :

- Les pages web étant sur le réseau Internet, une première étape consiste à les **trouver**. C'est le travail du *crawler*.
- Les pages sont aussi **nettoyées** et **indexées**, pour les transformer en une représentation permettant de recherche de l'information facilement à l'intérieur. C'est le processus d'*indexation*.
- Enfin, un mécanisme permet d'**interroger** les index créés et de renvoyer des documents à l'utilisateur. Une **fonction de similarité** permet de donner un score à un document en fonction de la requête.

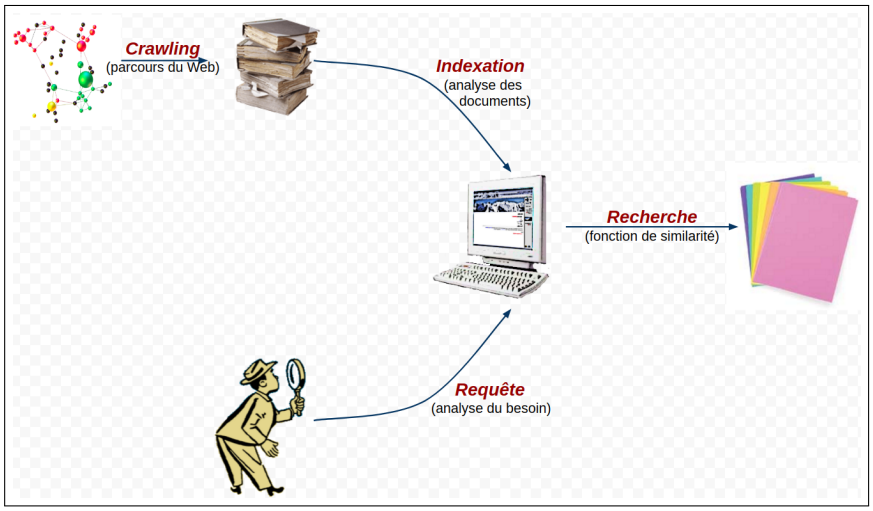

FIGURE 1 – Architecture générale d'un moteur de recherche.

Un bon moteur de recherche doit être à la fois **pertinent** (il répond bien au besoin d'information exprimé par l'utilisateur) et **efficace** (il répond relativement rapidement).

1. sauf que votre système ne pourra malheureusement pas interroger autant de pages web que ces géants.

Pour mener à bien ce projet, vous bénéficierez des aides suivantes :

— Les cours de recherche d'information en ligne de l'Université de Stanford [2] : des séries de courtes vidéos décrivant toutes les notions utiles à connaître. Nous ferons bien sûr le point sur ces cours ensemble.
— Du code python déjà fourni, ainsi que des exercices vous permettant de composer le squelette du moteur de recherche petit à petit.

2.1 Collection de pages web

Vous travaillerez tout d'abord sur une petite collection d'une centaine de textes, mais le projet final doit être réalisé sur un ensemble de 1,55 million de pages web provenant de sites de journaux de presse généraliste, en français. Cette collection vous sera fournie, ce qui pourra vous dispenser de l'étape du *crawling* ; cependant, vous pourrez réaliser un crawler de manière optionnelle, si vous le souhaitez.

2.2 Contraintes matérielles

Vous devrez vous assurer que votre système ne consomme pas plus de **1 Go de mémoire vive**, ni à l'indexation, ni à la recherche. D'autre part, votre index ne devra pas dépasser **60 % de la taille de la collection**. Le temps moyen de la réponse à une requête ne devra pas excéder **10 secondes**.

Ces contraintes vous obligeront à faire des choix en termes de représentation des informations, de structures de données, de fusion d'index. Tous ces choix devront bien sûr être décrits et justifiés dans le rapport.

2. https://www.youtube.com/watch?v=5L1qemKyUKA, du cours 18.1 à 19.8

3 Détails sur le rendu attendu

Cette section utilise un vocabulaire que vous ne maîtrisez probablement pas au moment de découvrir le sujet. Vous l'apprendrez bien vite !

3.1 L'indexation

Vous devrez employer **au moins deux techniques d'indexation** sur la collection : segmentation simple sans normalisation et avec *stemming*. Vous êtes libre de mettre en œuvre d'autres techniques si vous le souhaitez, et de choisir de supprimer les mots vides ou pas. Pour la pondération, la formule du **tf.idf** sera employée.

3.2 La recherche

Vous utiliserez le **modèle vectoriel** vu en cours et en TP, avec la **similarité cosinus**. Les requêtes seront de simples listes de mots-clés séparés par des espaces. Au moins deux versions du programme seront proposées : l'une s'appuyant sur l'index non normalisé, l'autre sur l'index avec *stemming*.

3.3 Évaluation

Une fois le moteur de recherche réalisé, il sera ensuite nécessaire d'évaluer sa pertinence et ses performances.

3.3.1 Pertinence

Vous évaluerez la pertinence des 20 premiers documents renvoyés par votre moteur de recherche, pour chacune des 10 requêtes ci-dessous. Vous définirez le protocole et les métriques d'évaluation vous-mêmes, avec l'aide de votre cours de recherche d'information (*"Modèles de recherche et évaluation"*, à partir de la page 48).

Les dix requêtes sont :

1. `Charlie Hedbo`
2. `volcan`
3. `playoffs NBA`
4. `accidents d'avion`
5. `laïcité`
6. `élections législatives`
7. `Sepp Blatter`
8. `budget de la défense`
9. `Galaxy S6`
10. `Kurdes`

Vous justifierez dans votre rapport le choix des métriques et la manière de juger la pertinence d'un document ; vous décrirez également les limites d'une telle méthode d'évaluation.

3.3.2 Performances

Vous fournirez toutes les mesures objectives pertinentes de la performance de votre système. Devront figurer **au minimum** :

— Le **temps de calcul** consacré à l'indexation (pour chaque index), et à la réponse moyenne à une requête.
— Vous comparerez également l'**espace disque** occupé par les différents index. Vous expliquerez les différences constatées.
— Enfin, vous estimerez la **mémoire occupée** lors de l'indexation et lors d'une requête. Vous fournirez une courbe montrant l'évolution de la consommation mémoire, en particulier pour l'indexation, par exemple avec l'aide de l'outil `jconsole`.