Projet Web Mining

Première partie

Pierre Senellart
pierre.senellart@telecom-paristech.fr

8 au 19 octobre 2012

Il s'agit de réaliser, en équipe et sur une période de deux semaines, un moteur de recherche sur les documents du Web. Le but est – en se restreignant éventuellement à un domaine précis, à un corpus de documents précis, en se concentrant sur une fonctionnalité précise, etc. – de « faire mieux que Google ».

La première étape (lundi 8 octobre) est d'établir un cahier des charges détaillé des fonctionnalités attendues, et de répartir les tâches entre les membre de l'équipe. Un debriefing aura lieu le mardi 9 octobre pour valider les choix avec l'enseignant et orienter le développement. Il y aura ensuite une réunion quotidienne pour faire le point sur l'avancement du projet.

Les deux premières semaines du projet sont consacrées à l'acquisition, l'extraction et l'indexation des données. Le projet continuera sur les trois semaines suivantes en se consacrant à la classification (clustering), catégorisation, réduction de dimensionalité, supervisées ou non, sur les données accumulées.

Le cahier des charges doit fournir des choix pour chacun des points suivants.

1 Domaine

On pourra décider de se restreindre, au choix, à :

— Une application donnée : sites d'actualités français, réseaux sociaux, annonces immobilières, sites en esperanto, etc.
— Une ou plusieurs hiérarchies de noms de domaines (p. ex., les sites des écoles de ParisTech, `lemonde.fr`, `inria.fr`, `legifrance.gouv.fr`, etc.)
— Des corpus ne nécessitant pas de crawl : Wikipedia (`http://dumps.wikimedia.org/`), crawls d'Internet Archive (p. ex., Wikileaks) disponibles sur `http://www.archive.org/details/web`, etc.

On peut aussi partir sur un crawl du Web dans son ensemble (évidemment restreint à quelques millions de pages vues les contraintes).

Quel que soit le choix retenu, il doit y avoir au minimum une centaine de milliers de documents indexés. Tenir compte des contraintes de politesse du crawl pour réfléchir à ce qu'il est raisonnable de crawler en deux semaines, si crawl il y a.

2 Fonctionnalités

Tout ou partie des fonctionnalités suivantes pourront être intégrées au système dans un premier temps :

— Crawler

— Détection de la langue
— Prétraitements linguistiques pour une langue donnée
— Indexation plein-texte
— Ranking (tf-idf, PageRank, etc.)
— Extraction d'informations
— Système de recommandation
— etc.

Chacune de ces fonctionnalités pourra soit être développée à partir de zéro, soit s'appuyer sur des outils existants, au choix. Dans le cas d'un développement à partir de zéro, prévoir [1] 1 jour-homme par fonctionnalité pour un comportement de base, quelques jours-homme pour quelque chose de plus abouti, de l'ordre d'une semaine-homme pour un produit fini pouvant s'intégrer avec d'autres composants, plus pour des fonctionnalités supplémentaires.

Le moteur devra également être accessible via une interface Web : une attention particulière sera portée à rendre cette interface dynamique et ergonomique ; la sécurité de l'interface Web sera également une priorité.

Il faudra également réfléchir aux fonctionnalités de classification, catégorisation et réduction de dimensionnalité qui pourront être intégrée dans la deuxième partie du projet ; au moins une partie des données produites devront être annotées pour permettre de faire tourner des algorithmes d'apprentissage supervisé (par exemple, un sous-ensemble de pages Web est annoté à la main avec une catégorie parmi n, et un algorithme de catégorisation apprendra à distinguer entre ces catégories).

3 Architecture et technologies

Une réflexion sera portée sur l'architecture, en fonction de la taille des données à traiter et des intérêts : architecture centralisée en utilisant un SGBD ou un moteur d'indexation existant, architecture centralisée avec stockage disque brut, architecture distribué avec HDFS et MapReduce, etc. Le pour et le contre de chaque technologie (langage de programmation, technologies Web, etc.) sera également débattu.

4 Travail en groupe

Vous pourrez (et devrez) utiliser pour le travail en groupe :
— Un Wiki disponible à l'URL `http://patrocle.enst.fr/~pesto/wiki/doku.php` pour le partage de la documentation ; des comptes peuvent librement y être créés.
— Le système de contrôle de version Git. L'URL du dépôt est ssh ://patrocle/home/pesto/pesto.git et les comptes Unix correspondants sont fournis lors de la première séance.

Une introduction à Git et au développement collaboratif sera donnée le mercredi 10 octobre.

5 Ressources disponibles

Une machine puissante et connectée en permanence est mise à disposition (patrocle.enst.fr, contactable en SSH et sur laquelle vous avez des comptes Unix) ; il est également possible d'utiliser les machines des salles de TP, et d'organiser ces dernières en cluster. Dans la mesure du raisonnable, d'autres ressources matérielles et logicielles peuvent également être mises à disposition.

1. Estimation grossière, certaines fonctionnalités étant plus simples que d'autres à implémenter.

6 Répartition des tâches

Chaque membre de l'équipe a un rôle bien défini en fonction de ses capacités et intérêts personnels. Le choix de ses rôles se fait conjointement et devra être inscrit, nominativement, dans le cahier des charges.

Les rôles suivant devront être représentés :

— 1 coordinateur ; son rôle est d'assurer la bonne répartition des tâches et le progrès du projet dans son ensemble, et de s'assurer que l'ensemble des composants développés peuvent interagir. Il est de sa responsabilité de décider de réorienter le développement de l'ensemble du projet en cas de besoin. Il peut mettre la main à la pâte pour aider sur chaque module.
— 1 ou plusieurs développeurs pour chaque module logiciel du moteur de recherche.
— 1 concepteur d'interface pour la partie « côté client » de l'interface Web.
— 1 programmeur Web pour la partie « côté serveur » de l'interface Web.
— 1 ou 2 responsables qualité, dont le rôle est d'une part de tester chacun des modules développés, d'autre part de mettre en place un protocole de comparaison des résultats du moteur avec ceux de moteurs existants (par exemple, Bing, Google) et de rendre compte au coordinateur et aux responsables du module correspondant. Ils auront également pour tâche de produire des données étiquetées qui serviront 1/ à réaliser des évaluations formelles de la qualité des modules 2/ à entraîner des algorithmes d'apprentissage.
— (éventuellement) Tout autre poste qui semblera nécessaire à l'exécution du projet.

7 Rendu du cahier des charges

Le cahier des charges doit être rédigé sur le Wiki, et il sera relu et validé avec l'ensemble du groupe le mardi 9 octobre au matin.