Accès et recherche d'information / Information retrieval

Composante
Polytech
Grenoble - INP,
UGA

- **Langue(s) d'enseignement:** Français
- **Ouvert aux étudiants en échange:** Oui
- **Code d'export Apogée:** KARI8M14

Présentation

Description

L'objectif des cours est de montrer les fondements scientifiques des tâches les plus répandues en Recherche d'Information (RI). Le souci principal est de proposer un exposé cohérent des algorithmes classiques développés dans ce domaine, et de connaître le mécanisme des outils de l'internet qu'on emploie tous les jours. Cette étude ne se limite pas à l'application initiale de RI et s'intéresse aussi aux problèmes connexes dans lesquels de nombreuses avancées techniques ont été réalisées ces dernières années.

1. Indexation, représentation et compression (2 séances)

Les constructions du dictionnaire et de l'index inversé, ainsi que la représentation vectorielle des documents, constituent le point de départ dans toutes manipulations et recherche en RI. Dans une collection de documents donnée, construire le dictionnaire ou le vocabulaire correspond à extraire une liste de termes utiles, caractéristiques des documents présents dans la collection. L'autre concept fondamental en RI est la constitution de l'index inversé. Il s'agit ici de construire, pour chaque terme du dictionnaire, la liste des index de documents contenant ce terme. Cette liste, aussi appelée liste inversée, rend l'appariement entre les requêtes et les documents de la collection plus efficace. Pour les très grandes collections de données, un problème majeur est le stockage de l'index et du dictionnaire dans la mémoire ou sur le disque. Le défi dans ces cas est de trouver un moyen de compression simple et rapide des données.

2. Recherche d'Information (3 séances)

Informations non contractuelles.
Dernière mise à jour le 22 octobre 2021

Ce chapitre constitue le cœur de ce module. Pour un besoin d'information donné, le système de recherche le transcrit sous forme d'une requête, constituée de mots-clés, et lorsque l'utilisateur regarde le résultat de la recherche, il voit les documents triés par ordre décroissant de pertinence. Si la requête est une expression booléenne, l'utilisation de l'index inversé permet de trouver facilement et en un temps minimal tous les documents qui satisfont cette requête. En revanche, les systèmes booléens purs ne permettent pas de retrouver les documents similaires au besoin d'information de l'utilisateur et ne contenant pas exactement les termes de la requête. Plusieurs modèles ont été développés pour pallier ce problème, depuis les modèles vectoriels jusqu'aux modèles probabilistes. De même, plusieurs stratégies, qui consistent à étendre la requête afin d'y inclure des termes similaires mais non mentionnés originellement par l'utilisateur, ont vues le jour afin d'enrichir ces différents modèles.

3. Recherche sur le web (1 séance)

La toile (ou le web) est un entrepôt dynamique et distribué de documents qui, par sa taille, par le manque de supervision dans la génération et la suppression de documents, ainsi que par la diversité du type de ces derniers, rend la recherche bien plus difficile que la recherche traditionnelle effectuée sur des collections classiques. Les premiers moteurs de recherche sur la toile reproduisaient néanmoins directement les méthodes de RI classiques, le défi principal étant de gérer des index inversés de très grandes tailles. La prise en compte, vers la fin des années 90, d'une des caractéristiques essentielle du web, à savoir les liens hypertexte reliant les documents entre eux, a permis, d'une part, de réaliser une meilleure indexation des pages web et, d'autre part, de donner un score de notoriété à chaque page sur la base de la topologie de la toile. Cela a conduit à la première génération des moteurs de recherche vraiment adaptés au web, dont Google fut le prototype. De nos jours, d'autres éléments sont pris en compte et les modèles utilisés reposent sur des techniques récentes d'apprentissage automatique.

4. Classification de documents. (4 séances)

Un système de classification de documents a pour but de catégoriser automatiquement une collection de documents suivant un ensemble de classes prédéfinies. Un exemple de tels systèmes est le catégoriseur de courriers électroniques incorporé dans la plupart des boîtes e-mails et qui place les courriers suspects automatiquement dans le dossier des courriers indésirables. Les systèmes de classification sont généralement conçus avec des techniques issues de l'apprentissage statistique et opèrent en deux phases. La première phase est la phase d'entraînement, lors de laquelle les paramètres du système sont réglés sur une base d'apprentissage contenant des documents avec leurs classes respectives. Durant cette phase le système apprend l'association entre les documents et leurs classes. C'est lors de la seconde phase, dite de test, que le système assigne une classe à chaque nouveau document entrant. Habituellement, les paramètres des systèmes d'apprentissage sont mis à jour périodiquement pendant le laps de temps où il n'y a pas de traitement à faire sur des documents arrivant.

In this course we introduce the scientifics fundamentals of the most important tasks in Information Retrieval.

1. Indexing, representation and compression (2 lectures)
2. Information Retrieval (3 lectures)
3. Information Retrieval on the web (1 lecture)
4. Document classification (4 lectures)

Informations non contractuelles.
Dernière mise à jour le 22 octobre 2021

Heures d'enseignement

Accès et recherche d'information / Information retrieval - CMTD	Cours magistral - Travaux dirigés	38h

Pré-requis recommandés

Notions de bases en probabilités

Basic notions in Probability

Période : Semestre 8

Évaluation initiale / Session principale - Épreuves

Libellé	Nature de l'enseignement	Type d'évaluation	Nature de l'épreuve	Durée (en minutes)	Nombre d'épreuves	Coefficient de l'épreuve	Remarques
						50/100	

Bibliographie

Cours basés sur le livre :
Modèles et Algorithmes en Recherche d'Information et ses Applications. Massih-Reza Amini et Eric Gaussier, 246 pages (avec une trentaine d'exercices corrigés), Éditions Eyrolles, Avril 2013, ISBN13 : 978-2-212-13532-9.

Infos pratiques

Lieu(x) ville

- Grenoble

Campus

- Grenoble - Saint-Martin d'Hères

Informations non contractuelles.
Dernière mise à jour le 22 octobre 2021