Institut
Mines-Télécom

Web mining

Pierre Senellart Stéphan Clémençon

université
PARIS-SACLAY

Table des matières

Internet vs. Web

2 choses bien différentes !

Internet Réseau d'ordinateurs interconnecté, organisé en sous-réseaux, avec protocoles d'adressage, de routage, et de transmission de l'information. Début des années 1970.

World Wide Web Système hypertexte reposant sur Internet, avec possibilité d'accéder à des ressources (textes, multimédia, contenu structuré) interliées et de soumettre des informations. Début des années 1990.

Contexte

- Web : inépuisable source d'informations
  - Centaines de milliards de pages Web
  - Millions de bases de données accessibles par le Web
- Nature très variée : texte, données structurées, textes semi-structurées, images, vidéos...
- Comment découvrir, indexer, fouiller, utiliser cette information ?
  - Veille
  - Aide à la décision
- Comment optimiser la diffusion de cette information ?
- Technologies des moteurs de recherche actuels et du futur ?

Table des matières

Technologies du Web

- Protocoles du World Wide Web
- Langages du Web (HTML, CSS, JavaScript. . .) vs CMS et Wikis
- Animation, Audio et Vidéo sur le Web
- Où en est la guerre des navigateurs ?
- Web et sécurité informatique
- Conception d'un crawler
- Calcul distribué sur le Web, architectures type MapReduce

Moteurs de recherche sur le Web

- Traitement informatique de la langue naturelle
- Recherche d'informations, bibliothèques électroniques
- Ranking, calculs efficaces
- Modèle du graphe du Web, typologie du graphe
- Scores d'importance sur le Web
- Contrer le *spamdexing*
- Réseaux sociaux
- Publicité ciblée, systèmes de recommandation

Le World Wide Web vu comme un graphe

Graphe orienté :

Nœuds : pages Web

Arêtes : hyperliens

Questions

Quelle structure ? Quels algorithmes ?

Caractéristiques du graphe du Web

Densité. Le graphe est-il peu dense ? ($|A| \ll |S|^2$) ?

Caractéristiques du graphe du Web

Densité. Le graphe est-il peu dense ? ($|A| \ll |S|^2$) ? Oui

Caractéristiques du graphe du Web

Densité. Le graphe est-il peu dense ? ($|A| \ll |S|^2$) ? Oui

Distance typique. Quelle est la distance moyenne entre deux nœuds ?

Caractéristiques du graphe du Web

Densité. Le graphe est-il peu dense ? ($|A| \ll |S|^2$) ? Oui

Distance typique. Quelle est la distance moyenne entre deux nœuds ?

Distance typique logarithmique

Caractéristiques du graphe du Web

Densité. Le graphe est-il peu dense ? ($|A| \ll |S|^2$) ? Oui

Distance typique. Quelle est la distance moyenne entre deux nœuds ?
Distance typique logarithmique

Transitivité. Si a est relié à b et c, est-ce que la probabilité que b est relié à c significativement plus grande que la probabilité que deux nœuds arbitraires soient reliés ?

Caractéristiques du graphe du Web

Densité. Le graphe est-il peu dense ? ($|A| \ll |S|^2$) ? Oui

Distance typique. Quelle est la distance moyenne entre deux nœuds ?
Distance typique logarithmique

Transitivité. Si a est relié à b et c, est-ce que la probabilité que b est relié à c significativement plus grande que la probabilité que deux nœuds arbitraires soient reliés ? Oui

Caractéristiques du graphe du Web

Densité. Le graphe est-il peu dense ? ($|A| \ll |S|^2$) ? Oui

Distance typique. Quelle est la distance moyenne entre deux nœuds ?
Distance typique logarithmique

Transitivité. Si a est relié à b et c, est-ce que la probabilité que b est relié à c significativement plus grande que la probabilité que deux nœuds arbitraires soient reliés ? Oui

Distribution des degrés. Quelle est la distribution des degrés des nœuds ?

Caractéristiques du graphe du Web

Densité. Le graphe est-il peu dense ? ($|A| \ll |S|^2$) ? Oui

Distance typique. Quelle est la distance moyenne entre deux nœuds ?

Distance typique logarithmique

Transitivité. Si a est relié à b et c, est-ce que la probabilité que b est relié à c significativement plus grande que la probabilité que deux nœuds arbitraires soient reliés ? Oui

Distribution des degrés. Quelle est la distribution des degrés des nœuds ?

Loi en puissance des degrés entrant et sortant

Pas seulement le graphe du Web !

- Mêmes caractéristiques dans :
  - les réseaux sociaux
  - les systèmes nerveux
  - les graphes d'interaction de protéines
  - les graphes de citations
  - etc.
- Majorité des cas : loi en puissance avec exposant entre -2 et -3 !
- Contre-exemples : graphes planaires, graphes de transports (plus de régularité, pas forcément de transitivité, plus grande distance typique, etc.).

Explications ? Conséquences sur la fouille de graphes ?

Table des matières

Apprentissage

- Apprentissage supervisé et non supervisé
- Catégorisation (classification)
- Classification (clustering)
- Extraction de structure
- Applications :
  - Reconnaissance de formes
  - Classement (ranking) optimal de documents
  - Prédiction
- Réduction de dimensionalité
- Évaluation statistique

Table des matières

Calendrier

Semaine 1 : Introduction au Web et aux moteurs de recherche (cours)

Semaines 2–3 : Acquisition de données du Web et conception d'interface (projet)

Semaines 4–6 : En parallèle, introduction à l'apprentissage (cours) et applications aux données acquises du Web (projet)

+ conférences d'ouverture (commune avec les autres PESTO basés à Télécom ?)

Projet

- Sujet au choix des étudiants
- Un seul sujet, travail en groupe avec répartition des tâches
- Comporte des aspects :
  - Acquisition de données du Web, Interfaces Web (première phase)
  - Apprentissage (deuxième phase)
- Exemples :
  - Moteur de recherche d'images
  - Comparateur de prix
  - Requêtes sémantiques sur les données du Web
  - Classification des politiques suivant leurs citations dans les médias (sujet 2011)
- Suivi régulier par les enseignants

Prérequis pour ce PESTO

- Bases de programmation (p. ex., cours Java de l'X)
- Volonté de travailler en groupe
- Possibilité de tâches non orientées vers la programmation : gestion de projet, études qualité (évaluation de la précision des résultats, etc.), rédaction technique, architecture, modélisation
- Aucun prérequis lié au Web, à l'apprentissage, à la recherche d'informations, à la fouille

Table des matières

Questions ?

pierre.senellart@telecom-paristech.fr
stephan.clemencon@telecom-paristech.fr