



Institut
Mines-Télécom



Web mining

Du texte au multim dia





Table des matières

Présentation générale

Web : technologies, moteurs de recherche
Apprentissage & multimédia

Organisation

Questions



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-structurés, 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

Présentation générale

Web : technologies, moteurs de recherche

Apprentissage & multimédia

Organisation

Questions



Technologies du Web

- Internet et le 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 :

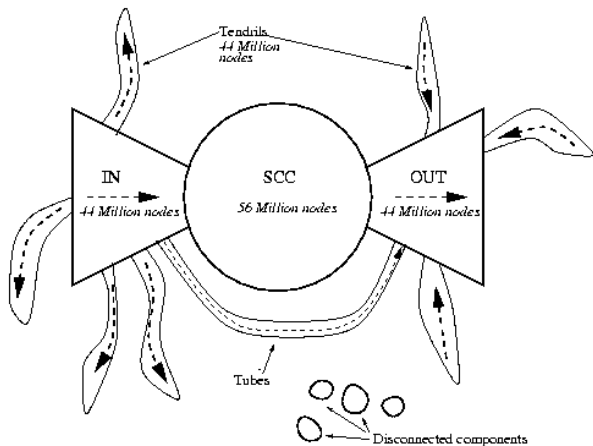
Nœuds : pages Web

Arêtes : hyperliens

Questions

Quelle structure ? Quels algorithmes ?

Structure du graphe du Web





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**

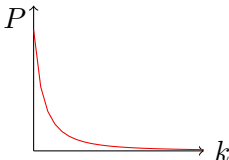
Distance typique. Quelle est la **distance moyenne** entre deux nœuds ?

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

Présentation générale

Web : technologies, moteurs de recherche

Apprentissage & multimédia

Organisation

Questions



Table des matières

Présentation générale

Web : technologies, moteurs de recherche
Apprentissage & multimédia

Organisation

Questions

Calendrier

		Lundi	Mardi	Mercredi	Jeudi	Vendredi
27/09-01/10		Cours	Cours	Conf./Visite	Cours	Cours
04/10-08/10		Cours	Cours	Conf./Visite	Cours	Cours
11/10-15/10		Cours	Cours	Conf./Visite	Cours	Cours
18/10-22/10	AM	Cours	Cours	Conf.	Cours	Cours
	PM	Projet	Projet	Projet	Projet	Projet
25/10-29/10		Projet	Projet	Projet	Projet	Projet
08/11-12/11	AM	Conf.	Conf.	Conf.		Conf.
	PM	Sout.	Sout.	Sout.		Sout.



Cours

Cours magistraux et travaux dirigés sur 4 sujets :

1. Technologies du Web (~20h)
2. Multimédia (~20h)
3. Recherche sur le Web (~20h)
4. Apprentissage et fouille de données (~20h)

Pierre SENELLART

Éric MOULINES

Pierre SENELLART

Éric MOULINES



Conférences d'ouverture

Sujets possibles des conférences d'ouverture :

- archivage du Web
- défis des moteurs de recherche actuels
- données et calcul distribué sur le Web
- protection des données sur le Web, watermarking
- utilisation des données du Web dans l'administration
- Web et régulation



Visites d'entreprises

En projet :

- Exalead
- projet Quaero
- Google Londres et Sony Londres
- start-ups de l'incubateur Télécom



Projet

Exemples de sujets :

- **Internet du futur** : vue d'ensemble, impact sur le contenu et l'accès au contenu
- **Web et propriété intellectuelle** : aspects légaux et aspects techniques
- **Moteur de recherche d'images sur le Web** : réalisation concrète



Table des matières

Présentation générale

Web : technologies, moteurs de recherche
Apprentissage & multimédia

Organisation

Questions



Questions ?

`pierre.senellart@telecom-paristech.fr`
`eric.moulines@telecom-paristech.fr`