

Moteurs de Recherche, Cours II

Cours Master 2, janvier 2010

Michel Habib
habib@liafa.jussieu.fr
<http://www.liafa.jussieu.fr/~habib>

janvier 2010

Plan

Actualités de la semaine

Graph Mining

Modèle économique d'un moteur de recherche

Les Algorithmes

Paradoxes de la sémantique

Sémantiques

Sémantiques

Energie

- ▶ Google = 2 Millions de machines qui tournent 24/24
- ▶ Energie nécessaire = 1 grosse centrale nucléaire
- ▶ Electricité = plus gros poste budgétaire de Google
- ▶ Google devient fournisseur d'électricité aux USA

- ▶ On peut chiffrer en énergie le coût d'une requête
- ▶ Domaine de recherche : réduire le coût d'une requête

Google en Chine

- ▶ Attaque sur gmail par le gouvernement chinois
- ▶ **Baidu** un moteur de recherche politiquement correct, possède un service inspiré de Wikipedia qui est interdit en chine.

- ▶ Déontologie !!!
Yahoo discrédité en chine depuis l'arrestation d'un opposant
- ▶ Quelle éthique pour notre domaine ?
- ▶ Comment oublier ? Twitter, Facebook

La fonction " Google suggests" a fait l'objet d'un procès perdu par Google.

La recherche de mots clés dans les pages permet :

- ▶ d'indexer les pages afin de retrouver l'ensemble des pages contenant un mot clé donné
- ▶ Catégoriser les pages (Science, Religion, Sport ...)

- ▶ Recherche par mots clés avec " " qui permet d'exprimer la proximité des mots
- ▶ Recherche avec des opérateurs logiques : AND, OR ou NOT ... (Possible avec Excite à vérifier)

Nouveau domaine de recherche **Graph Mining**
en français Fouille de Graphes.

Il s'agit d'extraire des connaissances à partir de gigantesques graphes (souvent dynamiques)
Une équipe ATT labs y travaille.

Techniques appliquées sur les Graphes de communications :
téléphone, mail
cela permet de :

- ▶ subscription fraud – new accounts with many fraudulent numbers in their calling circle are suspicious and generate alert
- ▶ targeted (viral) marketing – allows us to find clusters of customers who have high probability of taking a given product offer.
- ▶ repetitive debtors - delinquent customers who try and set up a new account are identified by their calling patterns and the new account can be shut down.

Le modèle économique

1. Accès gratuit pour un individu, payant pour une société
2. Affichage de bandeaux publicitaires
3. Liens publicitaires (paiement au clic)
4. Facturation de logs, et analyse de traces diverses (séquence de requêtes) et statistiques
5. Indexation immédiate de pages à la demande
6. Non indexation immédiate de pages à la demande!

Le modèle économique de Google est un peu différent, vu sa situation de monopole.

La fonction qui propose une nouvelle requête :
soit corrige l'orthographe d'un mot
soit cherche parmi les requêtes proches (à un mot près suivant la distance d'édition) celles qui a :

- ▶ le plus de réponses
- ▶ a été la plus demandée
- ▶ que vous avez le plus souvent demandée (ex : sncf.com)

Mais dans les deux cas, cela implique d'avoir une mémoire des requêtes déjà faites

Mais utilise-t-on cette mémoire pour répondre ? Précalcul des requêtes ou cache généralisé ?

PageRank :

dualité :

flot de matière ou interprétation probabiliste ?

Que mesure PageRank ?

- ▶ La pertinence d'une page
- ▶ cf. la paye chez Google

- ▶ Les limites des statistiques de mots clés (Rabelais et Dieu)
- ▶ La page recherchée ne contient pas nécessairement le mot de la requête (Page de Harvard, du MIT le mot est remplacé par un logo) ou les pages personnelles.
- ▶ Catégorisation bayésienne (donc heuristique car probabiliste)

Mot clés spéciaux par exemple :

Confidential do not distribute

permet de vérifier la stratégie de publication d'une société.

Les mots clés de la fin d'un texte ne sont pas référencés (vérifié sur Google)

Une typologie des requêtes

1. Savoir, connaissance : recherche d'information (48%)
2. Localisation : navigation (adresses, cartes, ...) (25%)
3. Achat en ligne (25 %)

Faut-il des moteurs de recherche spécialisés ?

Par exemple : Google Scholar pour le monde académique qui n'indexe que des textes.

A défaut une catégorisation des requêtes en fonction :

- ▶ du pays, de la langue
- ▶ des profils utilisateur
- ▶ de la requête elle-même, ce qui expliquerait l'absence de commutativité

Il faudrait utiliser :

Un moteur de recherche pour expert qui permette le paramétrage de la requête :

- ▶ sur la position des mots clés dans la page
- ▶ un paramétrage des occurrences du mot clé

Nécessité d'une approche expérimentale

Bien que le WEB soit une construction humaine munie d'un syntaxe (html ...)

personne n'en possède les plans.

Mais il possède une certaine sémantique qu'il s'agit de trouver à la manière des physiciens, sociologues en pratiquant des expérimentations.

- ▶ Il existe d'autres moteurs que Google
- ▶ Par exemple : Lucene un moteur écrit en java disponible (Logiciel libre) ...

Lucene

- ▶ Propose une boîte à outils pour construire son propre moteur de recherche. Idéal pour une application informatique comprenant un moteur écrit pour un domaine particulier : documentation d'une société, agence de voyage ...
- ▶ Lucene aide à spécifier l'indexation des documents
- ▶ Faut-il indexer les nombres ?
Mont blanc 4404 m
- ▶ Que faire du 4404 ?

Lucene

- ▶ Faut-il analyser les questions où les considérer comme des ensembles de mots clés ?
Les restaurants proches de chevaleret ?
- ▶ Avec quel ordre classer les réponses ?

Lucene

- ▶ En fonction de l'application on peut choisir l'indexation des documents (qui sont datés)
- ▶ L'analyse se faisant à partir de parsers spécialisés qui traitent du : texte, pdf, ps, html, msword, rtf, html, XML
- ▶ Mais aussi de la musique, des images, des vidéos

- ▶ Jeu Google : trouver une question en deux mots ayant ≤ 1 réponse.
(si possible sans guillemets dans la question)
- ▶ dorade droitière
poulpe ambitieuse ...
- ▶ Goolglewhack

- ▶ On retrouve tous les problèmes de l'analyse automatique des textes.
- ▶ Existe-t-il un formalisme universel ?
- ▶ IA
- ▶ Web sémantique

Représentation des connaissances

- ▶ Formalismes logiques :
nombreuses variantes logiques terminologiques ... logiques non-monotones, modales
algorithmique peu utilisable sur des grandes données
- ▶ Réseaux sémantiques, difficile d'y raisonner. Car les raisonnements diffèrent suivant la nature des liens.
Liens : d'héritage, d'instanciation, de composition ou de positionnement ...
- ▶ À base d'ontologies : arborescences de concepts
- ▶ Les graphes conceptuels un bon compromis

Graphes conceptuels

Graphe biparti construit sur deux types de sommets : concepts (instances) et relations (munies d'une arité)

Deux treillis structurent les concepts et les relations.

L'opération de base est le plongement de graphes qui préserve les types des noeuds.

Avantages

- ▶ Ce modèle généralise les ontologies.
- ▶ Il existe un modèle logique bien identifié
- ▶ un raisonnement algorithmiquement efficace.

Distance et proximité sémantiques

Il s'agit d'identifier qq concepts de bases (2000 par exemple)

Tout mot s'exprime alors comme un vecteur dans un espace de dimension 2000.

Et l'on peut parler de distance ou proximité entre mots, en calculant le produit scalaire de leurs vecteurs.

Ce modèle permet de gérer les polysémies des mots

Modèle très utilisé mais par très bien fondé scientifiquement.