Les algorithmes de classement utilisés dans les moteurs de recherche

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

Séminaire autour des objets numériques, 10 juin 2009, INRIA Rocquencourt

Plan

Plan

Plan

Plan

Plan

Plan

Introduction

Le Graphe du WEB

Pagerank

L'indexation des pages Web

Graph Mining

Quelques exemples de jeux algorithmiques

- ”Information is not Knowledge”, Albert Einstein

- "Information is not Knowledge", Albert Einstein
- "Information is not Knowledge. Knowledge comes from theory", W. Edward Deming

Fonctionnement d'un moteur de recherche

Données :une question (une chaîne de caractères)
Résultat : une liste ordonnée d'URL associées à la question.

Fonctionnement d'un moteur de recherche

Données :une question (une chaîne de caractères)
Résultat : une liste ordonnée d'URL associées à la question.

Comment cela marche

Fonctionnement d'un moteur de recherche

Données :une question (une chaîne de caractères)
Résultat : une liste ordonnée d'URL associées à la question.

Comment cela marche

1. Extraction du contenu de la question (i.e. quelques mots clés)

Fonctionnement d'un moteur de recherche

Données :une question (une chaîne de caractères)
Résultat : une liste ordonnée d'URL associées à la question.

Comment cela marche

1. Extraction du contenu de la question (i.e. quelques mots clés)
2. Recherche de toutes les pages WEB qui contiennent ces mots clés

Fonctionnement d'un moteur de recherche

Données :une question (une chaîne de caractères)
Résultat : une liste ordonnée d'URL associées à la question.

Comment cela marche

1. Extraction du contenu de la question (i.e. quelques mots clés)
2. Recherche de toutes les pages WEB qui contiennent ces mots clés
3. Tri et affichage d'une liste d'URL.

Tri des résultats

L'étape 3 est critique, car il peut y avoir plus de 100 000 réponses.

Tri des résultats

L'étape 3 est critique, car il peut y avoir plus de 100 000 réponses.

Une question très pertinente

Tri des résultats

L'étape 3 est critique, car il peut y avoir plus de 100 000 réponses.

Une question très pertinente

- Habib Terroriste

Tri des résultats

L'étape 3 est critique, car il peut y avoir plus de 100 000 réponses.

Une question très pertinente

- Habib Terroriste
- Google Results (February 2007) approx 504 000 for Habib terrorist. (0,11 seconds)

Un utilisateur normal ne lit que la première page des résultats
D'où l'absolue nécessité d'un bon classement des réponses

- Un moteur de recherche c'est :

- Un moteur de recherche c'est :
- un gigantesque graphe
  +

- Un moteur de recherche c'est :
- un gigantesque graphe
  +
- une gigantesque base de données
  +

- Un moteur de recherche c'est :
- un gigantesque graphe
  +
- une gigantesque base de données
  +
- des algorithmes efficaces

Nécessité du reverse engenieering car il y a beaucoup d'intox et de secret sur ce sujet

Nécessité du reverse engenieering car il y a beaucoup d'intox et de secret sur ce sujet

Ludique

Jeux expérimentaux avec les élèves sur la question "comment cela marche ?"

Eviter à tout prix la sémantique

En 1999, plusieurs chercheurs ont proposé une formulation récursive de l'importance d'une page.
Cette importance ne dépendant que de la structure des liens entre les pages html.

Eviter à tout prix la sémantique

En 1999, plusieurs chercheurs ont proposé une formulation récursive de l'importance d'une page.
Cette importance ne dépendant que de la structure des liens entre les pages html.

N'utiliser que la structure des hyperliens entre les pages permet d'éviter les analyses du contenu des pages (on évite ainsi le recours à des programmes d'analyse de la langue naturelle)

Interprétation des hyperliens

Méthode inspirée des études sur les citations entre scientifiques, par exemple le classement pondéré de G. Pinsky et F. Narin 1976

A cite B s'interprète comme A vote pour B

Interprétation des hyperliens

Méthode inspirée des études sur les citations entre scientifiques, par exemple le classement pondéré de G. Pinsky et F. Narin 1976
A cite B s'interprète comme A vote pour B

S. Brin, L. Page, R. Motwani, T. Winograd

Principe de l'algorithme de PageRank (Google)
Une page a un score d'autant plus élevé qu'elle est référencée par des pages ayant un score élevé.

Principe de la méthode Hits

J. Kleinberg

Pour chaque page on calcule de concert deux scores : un coefficient d'autorité et un coefficient d'annuaire (hub)
A page has a high hub score if it is references other pages with high authority scores. And a page has a high authority score if it is referenced by other pages with high hub scores

Le graphe du WEB

- Le graphe du Web un graphe orienté

Le graphe du WEB

- Le graphe du Web un graphe orienté
- Les sommets sont les pages html, appelées ici pages (10 billion de pages)

Le graphe du WEB

- Le graphe du Web un graphe orienté
- Les sommets sont les pages html, appelées ici pages (10 billion de pages)
- Les arcs correspondent aux hyperliens entre ces pages

Beaucoup de choses ont été écrites sur ce graphe . . .

Le fameux modèle du noeud papillon
Broder et al. (2000)

Beaucoup de choses ont été écrites sur ce graphe . . .

Le fameux modèle du noeud papillon
Broder et al. (2000)

Graphe petit monde

Les degrés vérifient une loi de puissance
$log(Prob(d^-(p) = k)) = \alpha - \lambda log(k)$ avec $\lambda = 2.1$ pour les degrés entrants et $\lambda = 2.72$ pour les degrés sortants.

Biais introduit par l'outil

T. Bennouas, F. de Montgolfier 2007

La plupart des propriétés trouvées proviennent en fait des méthodes choisies pour l'exploration du graphe

Biais introduit par l'outil

T. Bennouas, F. de Montgolfier 2007

La plupart des propriétés trouvées proviennent en fait des méthodes choisies pour l'exploration du graphe

BFS

Un parcours en largeur explique le modèle du noeud papillon.

Exploration

L'exploration du graphe du Web est un problème techniquement difficile d'informatique distribuée (des programmes appelés robots suivent les liens)

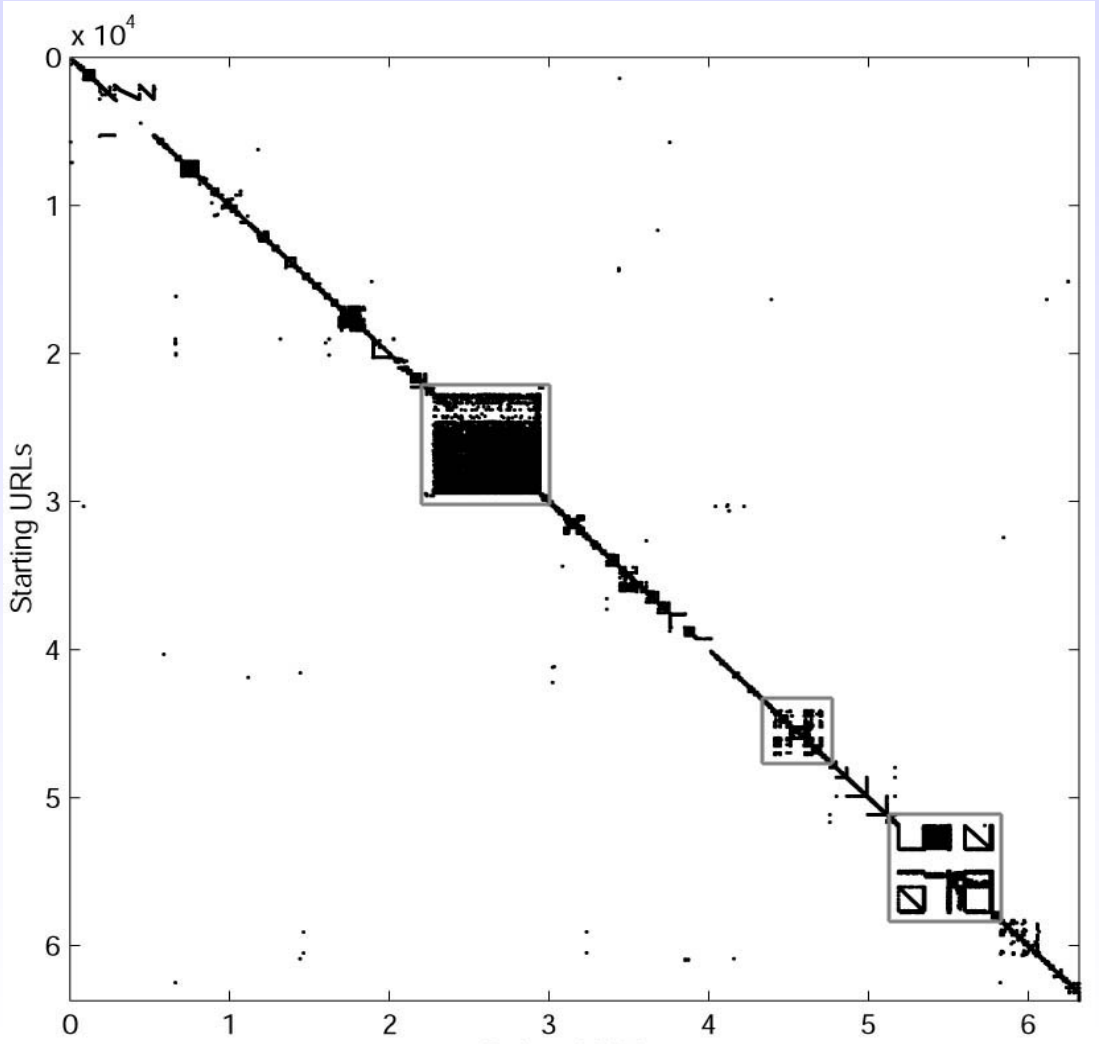
Exploration

L'exploration du graphe du Web est un problème techniquement difficile d'informatique distribuée (des programmes appelés robots suivent les liens)

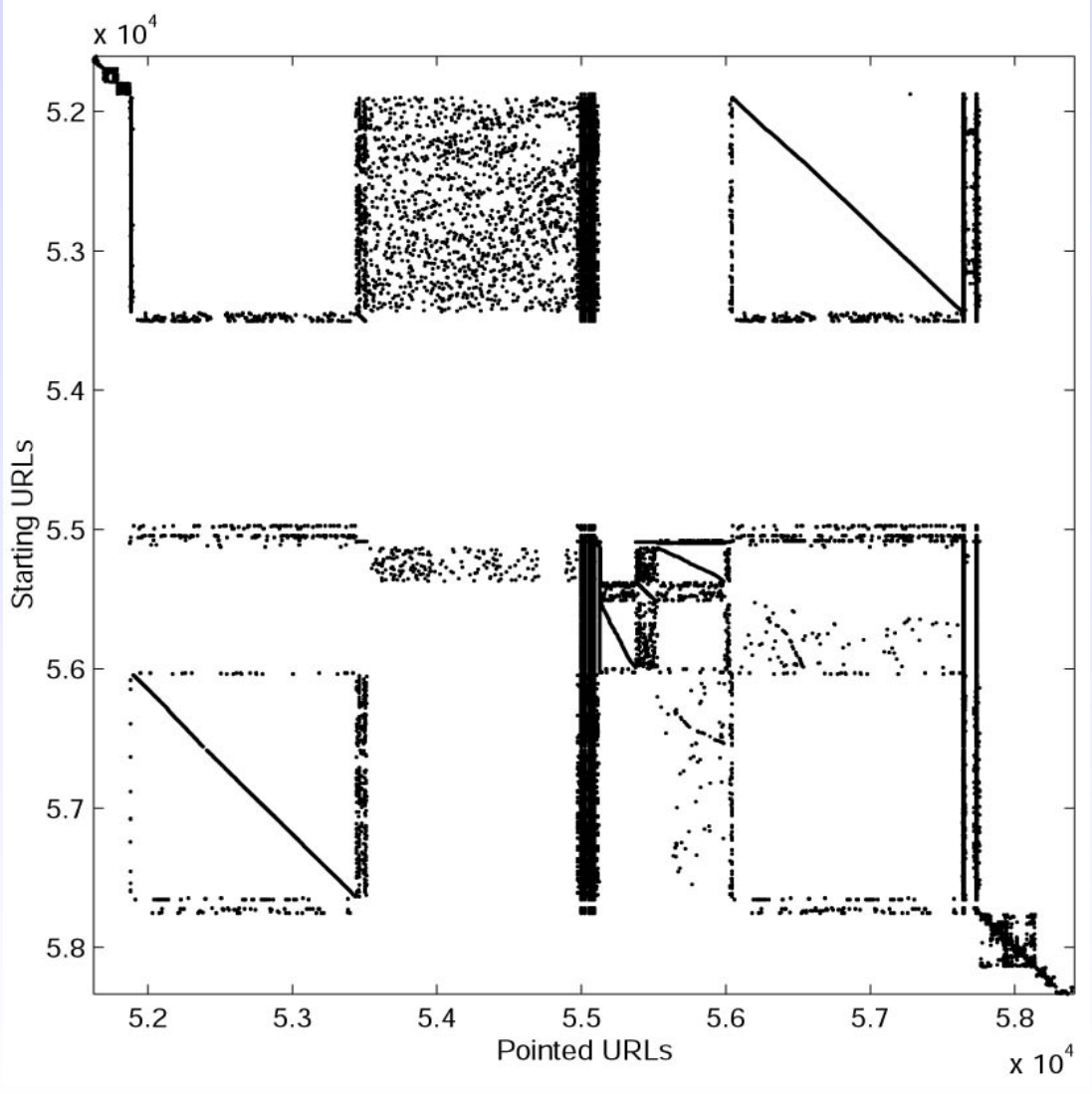
Graphes du Web

P. Boldi et son groupe de recherche propose des données publiques, ainsi qu'un logiciel BV graphs qui permet de compresser les graphes du Web avec 2-3 bits peof URL names

Matrice ordonnée par l'ordre alphabétique des noms des URL



Un zoom autour de la diagonale



- Un graphe peu dense représentable très efficacement

- Un graphe peu dense représentable très efficacement
- Algorithmes parallélisables

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.

Un modèle linéaire

Une sorte de flot

Soit $R_n(p)$ le coefficient PageRank de la page p à l'étape n du calcul et soit $R_{n+1}(p,q)$ la quantité qui traverse l'arc pq entre les étapes n et $n+1$.

Un modèle linéaire

Une sorte de flot

Soit $R_n(p)$ le coefficient PageRank de la page p à l'étape n du calcul et soit $R_{n+1}(p,q)$ la quantité qui traverse l'arc pq entre les étapes n et $n+1$.

L'équation

$$R_{n+1}(p) = \Sigma_{qp} R_{n+1}(q,p)$$

- Avec l'hypothèse de l'équirépartion du coefficient sur les liens sortant d'une page q

- Avec l'hypothèse de l'équirépartion du coefficient sur les liens sortant d'une page q
- On obtient
  $R_{n+1}(q,p) = \frac{1}{degre(q)} R_n(q)$
  pour tout arc qp.

- Avec l'hypothèse de l'équirépartion du coefficient sur les liens sortant d'une page q
- On obtient
  $R_{n+1}(q,p) = \frac{1}{degre(q)} R_n(q)$
  pour tout arc qp.
- D'où
  $R_{n+1}(p) = \Sigma_{qp} \frac{1}{degre(q)} R_n(q)$

Vectoriellement

- $R_{n+1} = A^T R_n$ où A est une sorte de matrice d' incidence du graphe du Web.

Vectoriellement

- $R_{n+1} = A^T R_n$ où A est une sorte de matrice d' incidence du graphe du Web.
- $A[p, q] = \frac{1}{d+(p)}$ si pq est un arc et 0 sinon

Vectoriellement

- $R_{n+1} = A^T R_n$ où A est une sorte de matrice d' incidence du graphe du Web.
- $A[p, q] = \frac{1}{d+(p)}$ si pq est un arc et 0 sinon
- Quand la suite R_n converge, sa limite est le vecteur propre associé à la valeur propre 1.

Pourquoi PageRank est-il tant utilisé ?

1. Convergence très rapide

Pourquoi PageRank est-il tant utilisé ?

1. Convergence très rapide
2. Le calcul peut se faire ligne à ligne en utilisant un codage compact du graphe

Pourquoi PageRank est-il tant utilisé ?

1. Convergence très rapide
2. Le calcul peut se faire ligne à ligne en utilisant un codage compact du graphe
3. Le calcul se parallélise simplement.

Pourquoi PageRank est-il tant utilisé ?

1. Convergence très rapide
2. Le calcul peut se faire ligne à ligne en utilisant un codage compact du graphe
3. Le calcul se parallélise simplement.
4. Il y a plusieurs interprétations mathématiques intéressantes du calcul

Convergence

A l'aide du théorème de Perron Froebenius

La convergence est assurée si le graphe est fortement connexe et si le pgcd des longueurs des circuits est 1.
Ce qui est impossible à vérifier sur le graphe du Web. Plusieurs astuces sont utilisées pour assurer la convergence du calcul.

Interprétation à l'aide des chaînes de Markov

A est une matrice stochastique et la limite de $R_n(p)$ peut se comprendre comme la probabilité qu'un surfeur aléatoire visite la page p.

Le vecteur R final n'est rien d'autre que la distribution stationnaire d'une marche aléatoire sur le graphe du Web

Un effet de bord ?

M. Bouklit et F. Mathieu ont essayé de modéliser plus avant le comportement d'un surfeur en introduisant par exemple la touche Retour (undo)
Le classement obtenu n'avait pas l'air significativement meilleur.
PageRank est-il un flot de matière ou une probabilité ?

Proposition de projet avec les élèves

Mise au point d'une programmation (ou du fonctionnement à la main) sur des petits exemples de Pagerank (genre séance d'exercices TP ou TD)
Vecteur initial
La question de la forte connexité

Le facteur ZAP

On initialise à $1/N$ le coefficient de PageRank de toutes les pages où N est le nombre total de pages du graphe du Web
$R_{n+1}(p) = \frac{d}{N} + (1-d).\Sigma_{qp}\frac{1}{degre(q)}R_n(q)$
on propose de prendre d le facteur ZAP entre $0.1 et 0.2$

Retour sur le fonctionement d'un moteur de recherche

1. Précalcul d'un fichier inversé des Pages Web, dans une gigantesque Base de données distribuée

Retour sur le fonctionement d'un moteur de recherche

1. Précalcul d'un fichier inversé des Pages Web, dans une gigantesque Base de données distribuée
2. Extraction du contenu de la question (i.e. quelques mots clés)

Retour sur le fonctionement d'un moteur de recherche

1. Précalcul d'un fichier inversé des Pages Web, dans une gigantesque Base de données distribuée
2. Extraction du contenu de la question (i.e. quelques mots clés)
3. Recherche de toutes les pages WEB qui contiennent ces mots clés et calcul d'un score pondéré pour chaque page (le score dépend des mots clés de la question)

Retour sur le fonctionement d'un moteur de recherche

1. Précalcul d'un fichier inversé des Pages Web, dans une gigantesque Base de données distribuée
2. Extraction du contenu de la question (i.e. quelques mots clés)
3. Recherche de toutes les pages WEB qui contiennent ces mots clés et calcul d'un score pondéré pour chaque page (le score dépend des mots clés de la question)
4. Filtrer les pages résultats à l'aide d'un profil d'utilisateur.

Retour sur le fonctionement d'un moteur de recherche

1. Précalcul d'un fichier inversé des Pages Web, dans une gigantesque Base de données distribuée
2. Extraction du contenu de la question (i.e. quelques mots clés)
3. Recherche de toutes les pages WEB qui contiennent ces mots clés et calcul d'un score pondéré pour chaque page (le score dépend des mots clés de la question)
4. Filtrer les pages résultats à l'aide d'un profil d'utilisateur.
5. Trier les pages obtenues à l'aide de PageRank (et de quelques petites astuces secrètes) et afficher cette liste ordonnée d'URL.

Calcul de score

Score pondéré

Construit à partir de :

Calcul de score

Score pondéré

Construit à partir de :

1. Les mots apparaissent dans le titre de la page (ou le chemin d'accès)
   (exemple French Military Victories)

Calcul de score

Score pondéré

Construit à partir de :

1. Les mots apparaissent dans le titre de la page (ou le chemin d'accès)
   (exemple French Military Victories)
2. Les mots apparaissent α fois dans la description de l'entête de la page.

Calcul de score

Score pondéré

Construit à partir de :

1. Les mots apparaissent dans le titre de la page (ou le chemin d'accès)
   (exemple French Military Victories)
2. Les mots apparaissent α fois dans la description de l'entête de la page.
3. Les mots apparaissent β fois dans la page, avec préférence pour le début de la page.

Un peu de technologie

La base de données est distribuée sur des centaines de milliers de machines (PCs ou serveurs)
Grosse consommation d'énergie électrique
Vu le taux de panne, chaque jour plusieurs centaines de machines sont en panne
Nécessité d'une grande redondance des informations
Comment faire ?

Les bombes de Google

- Le nom associé à la balise html d'une page q qui pointe sur une page p est utilisé pour indexer la page p

Les bombes de Google

- Le nom associé à la balise html d'une page q qui pointe sur une page p est utilisé pour indexer la page p
- Cela permet de fabriquer des bombes

Les bombes les plus célèbres

1. Talentless hacker versus Andy Pressman made by Adam Mathes in 2001.

Les bombes les plus célèbres

1. Talentless hacker versus Andy Pressman made by Adam Mathes in 2001.
2. Miserable failure versus George W. Bush made by George Johnston 2003

Les bombes les plus célèbres

1. Talentless hacker versus Andy Pressman made by Adam Mathes in 2001.
2. Miserable failure versus George W. Bush made by George Johnston 2003
3. Sarkozy versus Iznogood

Les bombes les plus célèbres

1. Talentless hacker versus Andy Pressman made by Adam Mathes in 2001.
2. Miserable failure versus George W. Bush made by George Johnston 2003
3. Sarkozy versus Iznogood
4. Ministre Blanchisseur versus Renaud Donnadieu de Vabres

Les bombes les plus célèbres

1. Talentless hacker versus Andy Pressman made by Adam Mathes in 2001.
2. Miserable failure versus George W. Bush made by George Johnston 2003
3. Sarkozy versus Iznogood
4. Ministre Blanchisseur versus Renaud Donnadieu de Vabres
5. ...

Dans la plupart des cas le mot de la balise est un bon mot clé pour une page, et une grande partie du succès de Google vient de cela.

Dans la plupart des cas le mot de la balise est un bon mot clé pour une page, et une grande partie du succès de Google vient de cela.

First Google answer was : We just show what is between the Web pages.

Dans la plupart des cas le mot de la balise est un bon mot clé pour une page, et une grande partie du succès de Google vient de cela.

First Google answer was : We just show what is between the Web pages.

Une bombe explose uniquement après la mise à jour des coefficient de PageRank (un mois)
Mais après elle résiste au temps, il est difficile de s'en débarasser !

Existe-t-il une solution algorithmique pour la détection des bombes ?

01/26/2007

Google announced today a modification to their search algorithm that minimizes well-known googlebombing exploits. Searches on "miserable failure" and their ilk no longer bring up political targets. The Google blogger writes : By improving our analysis of the link structure of the web, Google has begun minimizing the impact of many Googlebombs. Now we will typically return commentary, discussions, and articles about the Googlebombs instead.

L'analyse du graphe du Web permet :

- de calculer un ordre total sur les pages PageRank

L'analyse du graphe du Web permet :

- de calculer un ordre total sur les pages PageRank
- de calculer d'autres ordres (Annuaire, Autorité, Spam)

L'analyse du graphe du Web permet :

- de calculer un ordre total sur les pages PageRank
- de calculer d'autres ordres (Annuaire, Autorité, Spam)
- de rechercher des communautés entre pages fortement reliées

L'analyse du graphe du Web permet :

- de calculer un ordre total sur les pages PageRank
- de calculer d'autres ordres (Annuaire, Autorité, Spam)
- de rechercher des communautés entre pages fortement reliées
- d'analyser les logs (traces des visites)

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

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 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 verifier)

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

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

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.

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

Le modèle économique

1. Accès gratuit pour un individu, payant pour une société
2. Affichage de bandeaux publicitaires

Le modèle économique

1. Accès gratuit pour un individu, payant pour une société
2. Affichage de bandeaux publicitaires
3. Liens publicitaires (paiement au clic)

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

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

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 ! Si on veut faire disparaître une page compromettante

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 ! Si on veut faire disparaître une page compromettante
7. Moins correct
   Affichage dans les résultats non-publicitaires dans la première page de résultats moyennant finance (paiement au clic, i.e. on paye pour 2000 clics) pour certaines requêtes.

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

Commutativité

La commutativité des mots clés dans une requête ?
Exemples les réponses ne sont pas classées dans le même ordre si l'on pose les questions :
Nicolas Sarkozy
Sarkozy Nicolas
ou encore Nicolas Sarkozi
ou Sarko

Ni Altavista, ni Exalead, ni Google, ni Ask ne sont commutatifs ! [1]
Comment l'expliquer ? Par une catégorisation des noms : prénom versus nom de famille ?

[1] Il existe plusieurs milliers de moteurs de recherche, je n'ai pas tout essayé

Trouver des questions ayant peu de réponses différentes

- Jeu Google : trouver une question en deux mots ayant ≤ 1 réponse.
  (si possible sans guillemets dans la question)

Trouver des questions ayant peu de réponses différentes

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

Trouver des questions ayant peu de réponses différentes

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

- Les limites des statistiques de mots clés (Rabelais et Dieu)

- 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 remplaçé par un logo) ou les pages personnelles.

- 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 remplaçé 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.

Une typologie des requêtes

1. Savoir, connaissance : recherche d'information (48%)

Une typologie des requêtes

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

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

A défaut une catégorisation des requêtes en fonction :

- du pays, de la langue
- des profils utilisateur

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 paramètrage de la requête :

- sur la position des mots clés dans la page

Il faudrait utiliser :

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

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

Quelques références

- T. Bennouas, PhD Thesis, Montpellier University, 2005.
- M. Bouklit, PhD Thesis, Montpellier University, 2006.
- S. Brin, L. Page, R. Motwani, T. Winograd, The PageRank citation ranking : bringing an order to the Web, Technical Report 1999-0120, Computer Science Dept. Standford, 1999.
- J. Kleinberg, Authoritative sources in a hyperlinked environment, J. of the ACM, 1999.
- A.N. Langville, C.D. Meyer, Google ?s PageRank and beyond, Princeton University Press, 2006.
- F. Mathieu, PhD Thesis, Montpellier University, 2004.

Merci de votre attention ! !