Méthodes automatiques pour la détection du spamdexing

Orange Labs

Tanguy Urvoy, Recherche & Développement
26 mars 2009, DAPA

diffusion libre

orange™

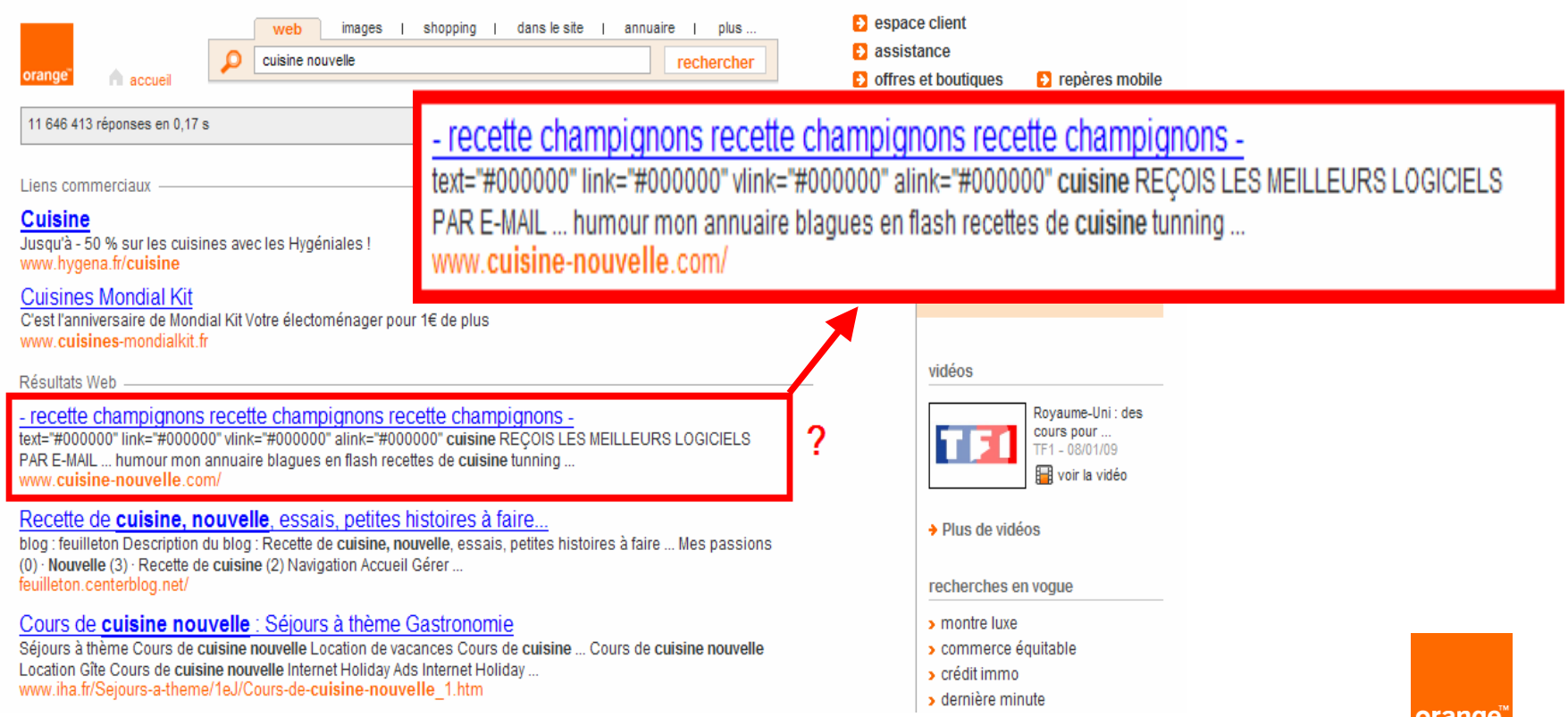
Le projet ANR MADSPAM

- Méthodes Automatiques pour la Détection de SPAMdexing sur les Grands Réseaux d'Information
- Site Web : http://madspam.org
- Objectifs :
- Développement d'outils génériques et pérennes de lutte contre le spamdexing
- Mise en œuvre sur deux plateformes :
  - Hébergeur de blogs (Blogspirit)
  - moteur (Orange)

Spamdexing

Définition

- different du spam e-mail
- techniques permettant d'améliorer artificiellement le classement d'un site sur les moteurs Web



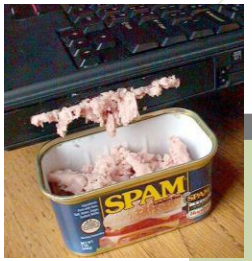
Ferme à liens, cuisine et illusions

- le site qui parle (entre autre) de champignons : cuisine-nouvelle.com
- fortement connecté avec des sites similaires comme mes-illusions.com
- ces sites n'offrent pas de contenu réel :
  - ils sont générés de manière automatique
  - c'est une **ferme à liens** ventilée sur plusieurs milliers de domaines.

Camouflage javascript, 8-gramme caractéristique

Les effets du spamdexing

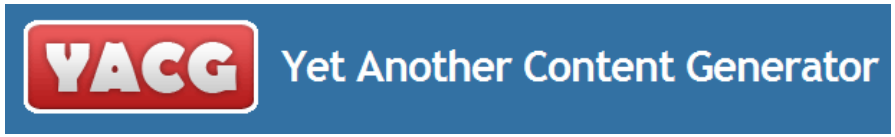
- pour l'utilisateur
  - le spam empêche d'accéder aux informations recherchée
  - énerve en proposant des publicités non sollicitées

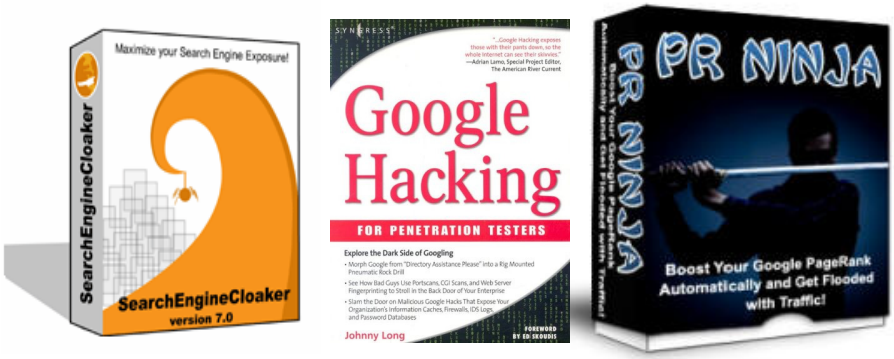


- pour les hébergeurs
  - gaspille de la bande passante
  - pollue les bases
  - dégrade la qualité du service
- pour les moteurs
  - gaspille de la bande passante lors du crawl
  - pollue les bases et les index
  - dégrade la pertinence des listes réponse

Taxonomie du spamdexing

- contenu
  - bourrage de mots clés
  - plagiat, Web-scraping, feed-scraping
  - générateurs de contenu :
    - pipotrons
    - patchworks, générateurs markoviens
- liens
  - pages satellites, échanges, fermes à liens
  - "pompes à pagerank"
- camouflage
  - hacks : blanc/blanc, cloaking, scripts, feuilles de style…
  - abondance de signaux faibles
- parasitage
  - recyclage de domaines expirés, cybersquatting
  - pollution ou piratage de sites réputés fiables
    - **blogs**, forums, petites annonces…
  - botnets
- variantes
  - MFA, clickbots (ClickBot.A)
  - fishing
  - générateurs d'amis



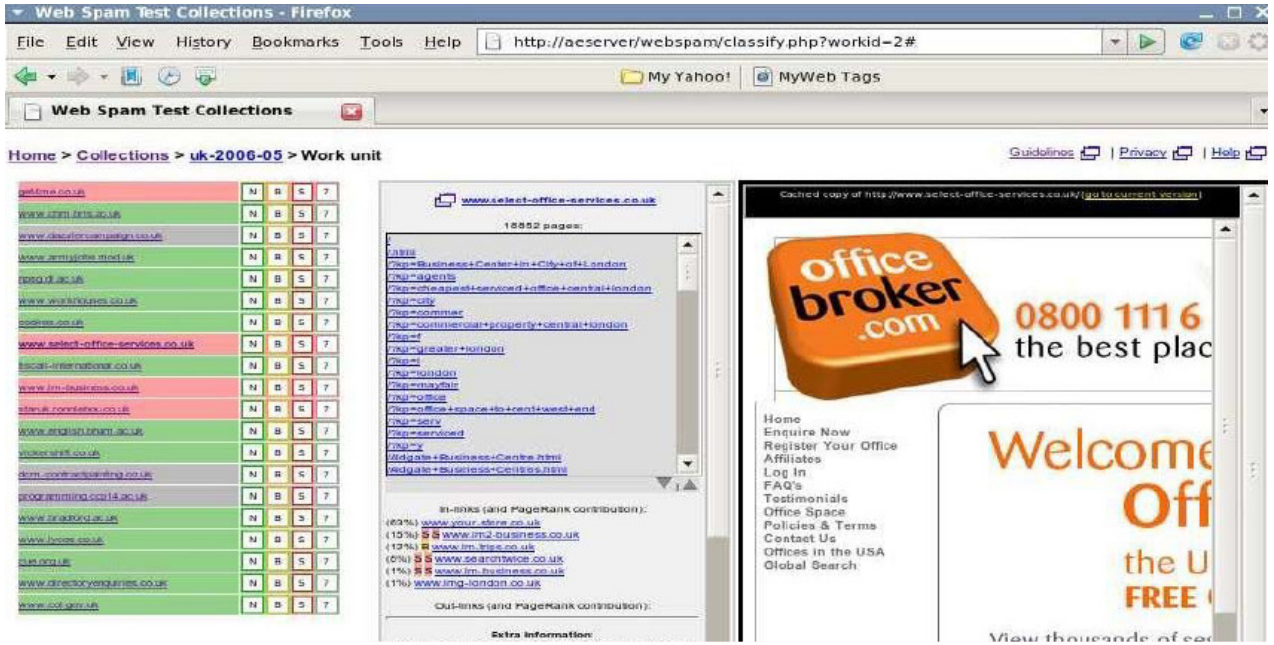


Méthodes de détection

Remarques générales sur le spamdexing

- redondance
  - production automatique d'un volume important de texte
  - plagiat, self-plagiat
- graphe des liens
  - le spam est pointé par du spam
  - le spam pointe aussi sur du ham
- aspect adverse
  - complexité croissante
  - les méthodes ad-hoc se dégradent vite (pour les spammeurs aussi)
  - **les méthodes statistiques sont les plus les adaptées** (pour les spammeurs aussi)
- sources de données fiables pour l'apprentissage et le test ?
  - équipe éditoriale, règles expertes
  - panels d'utilisateurs
  - campagnes d'étiquetage (c.f. Web Spam UK-2006 et UK-2007)

Corpus WebSpam UK-2006 et UK-2007



- agrément modéré entre les juges (Sur UK-2006 : Kappa=0.56)
  - impact des liens : amas de sites "borderline" = spam ? site cible = spam ?
  - subjectivité : qualité vs spamicité ?

- UK-2006 : 26% de spam, UK-2007 6% de spam
  - réelle diminution du spam?
  - camouflage plus performant, meilleure stratégie de crawl ?

Niveaux de filtrage antispamdexing

- crawl/indexation : empilement de classifieurs spam/unknown/ham
  - filtrages "low-cost" selon IP,URLs, selon contenu individuel
  - filtrages avancés :
    - ad-hoc : simulateur de navigateur, crawl furtif
    - anti-markoviens : thèse Thomas Lavergne le 3 avril à l' ENST
    - doublons, plagiat, stylométrie HTML : signatures
    - analyse des liens, des similarités : propagation d'étiquettes
    - équipe éditoriale
  - index spécialisés

- ranking
  - durcissement des algorithmes
  - semi-supervisé (trustrank like)
  - détection des requêtes "à risque"

stylométrie HTML

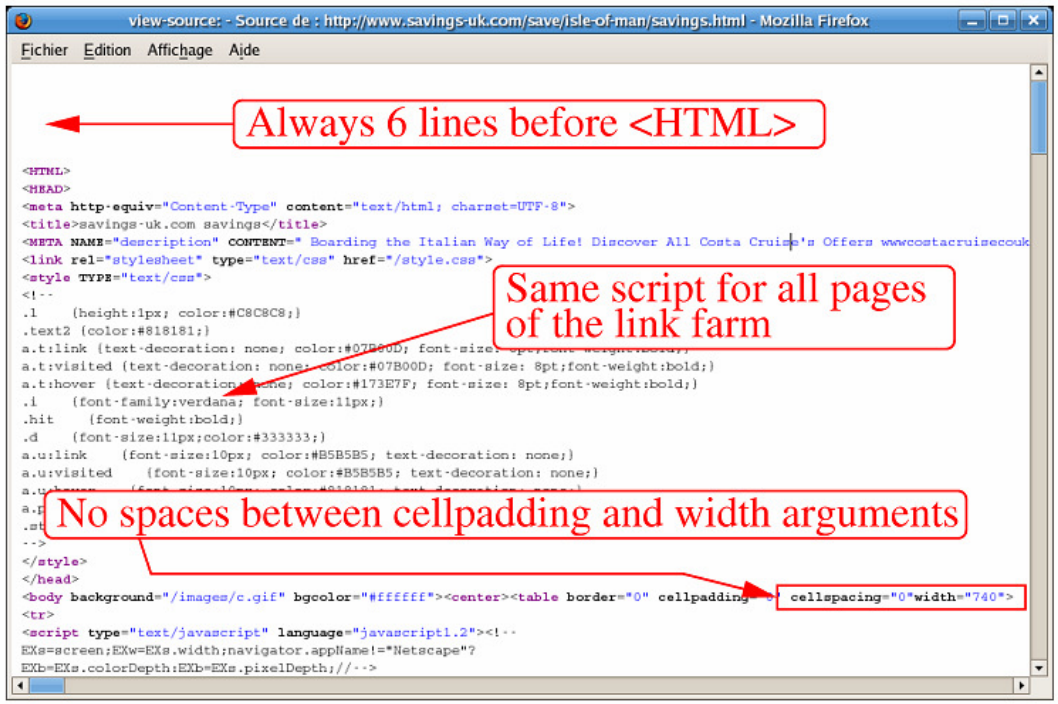
- les scripts perl manquent de créativité

get4me.co.uk (index)

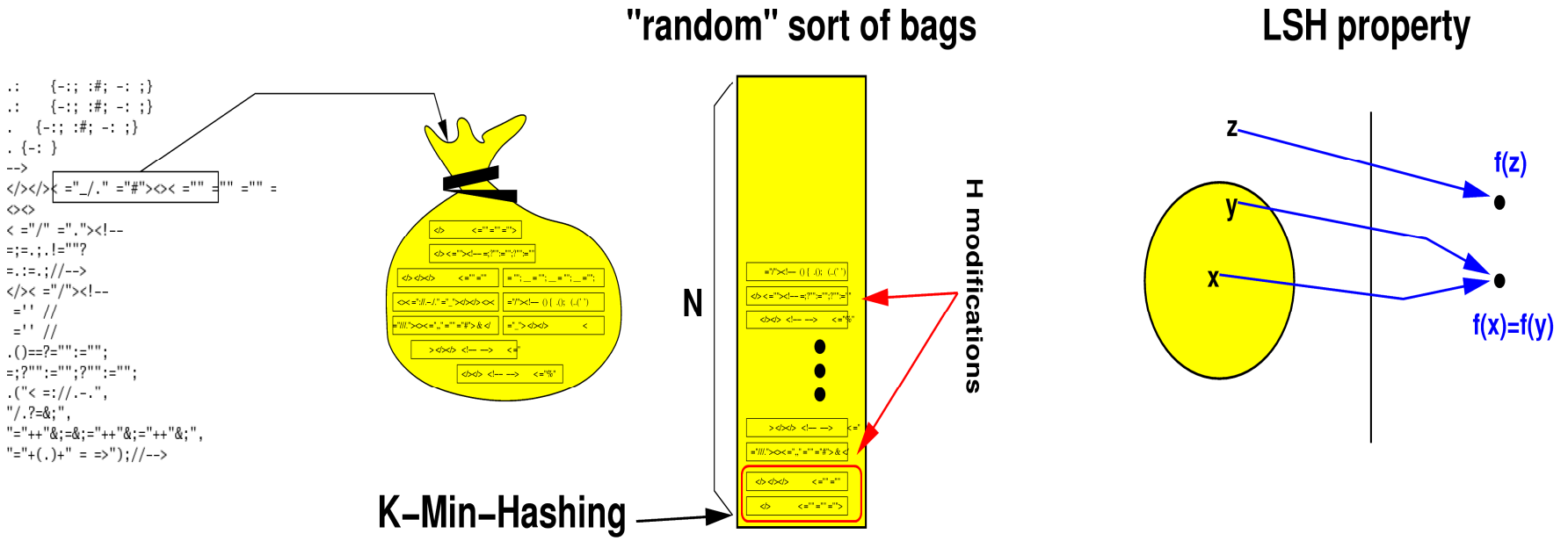
www.goto4me.co.uk

www.injuryclaim-uk.com

4u-insurance.co.uk

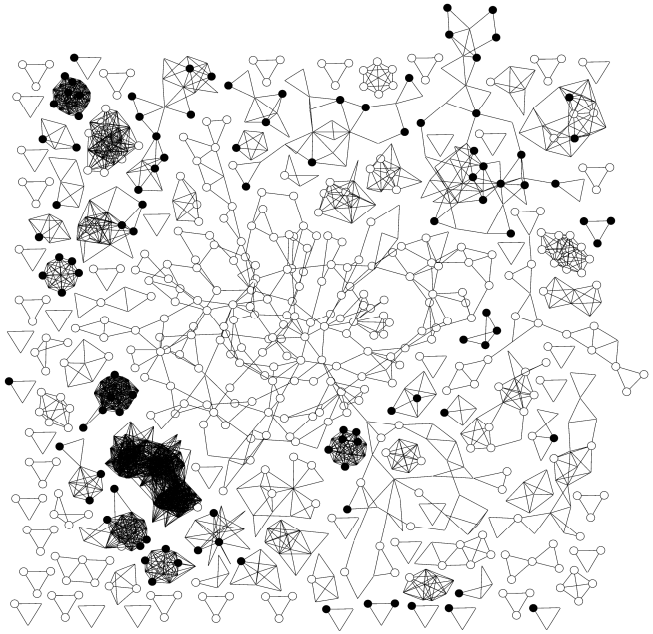


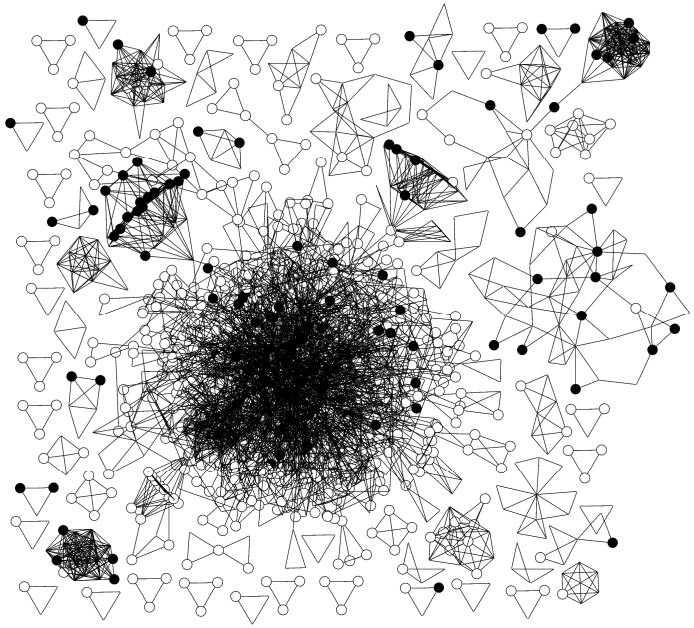
Calcul de signatures "de style"



1. extraction d'un squelette HTML
2. modélisation en sac de séquences
3. calcul de signatures LSH 64/128 bits par minhashing
4. clustering sur toute la base ou filtrage ppv

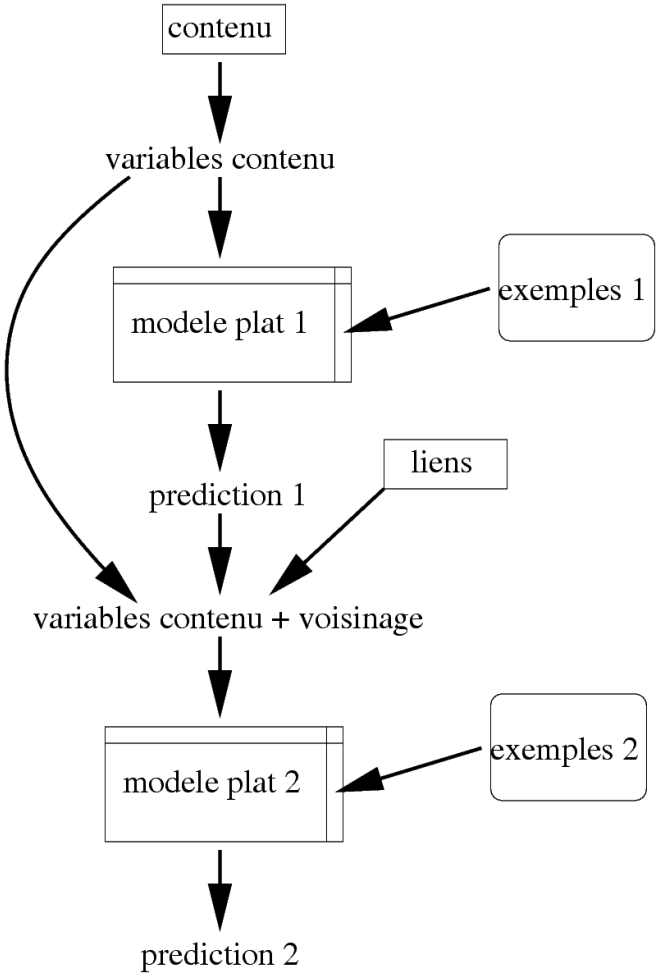
Propagation d'étiquettes : analyse liens/contenus



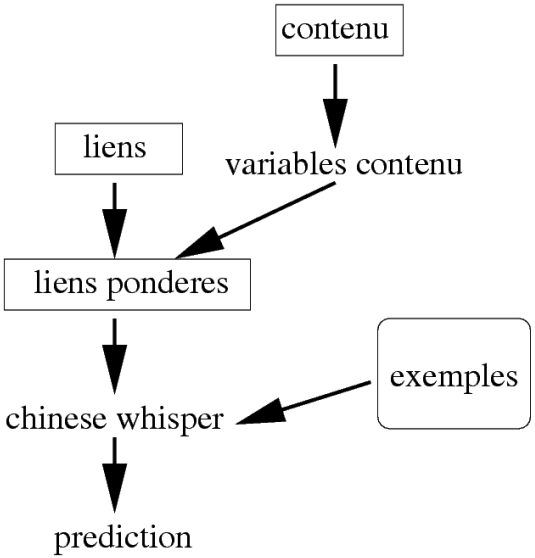


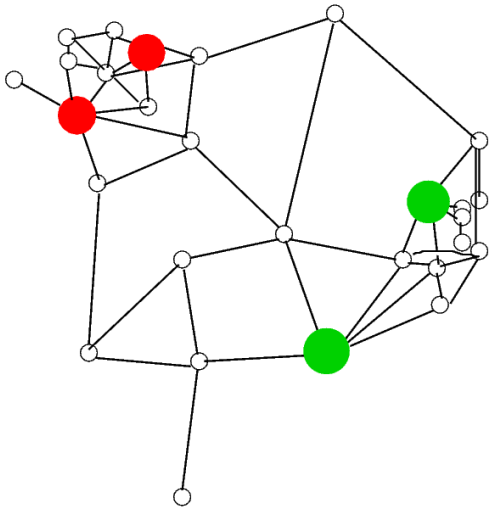
- graphes partiellement étiquetés + contenus
- nombreuses méthodes
  - renforcement du diagnostic contenu par voisinage
  - pondération des arcs selon contenu et propagation
- workshop graphlab 2007 et challenges WebSpam

Propagation d'étiquettes : analyse liens/contenus



- **stacked graphical learning** : renforcement du diagnostic contenu par voisinage
- **chinese whisper** : pondération des arcs et propagation jusqu'au point fixe





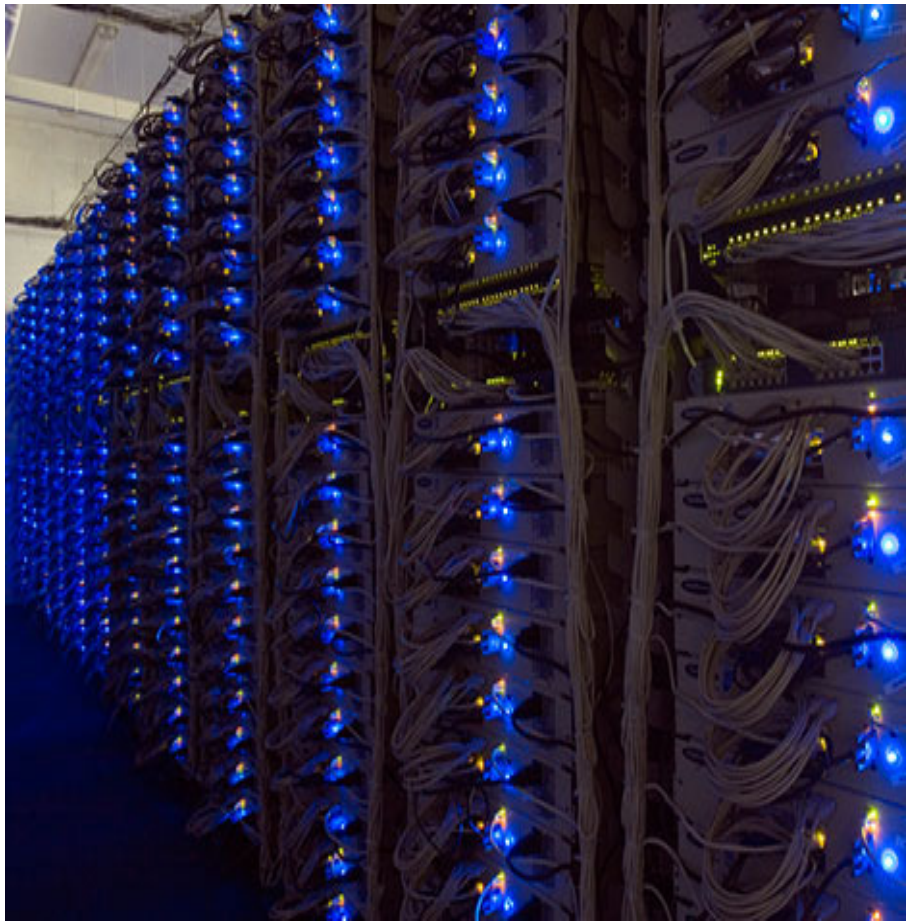
merci

diffusion libre

orange™

Annexes

Apprentissage adverse et sensibilité aux coûts



- un modèle statistique contre un autre
  - challenges adverses ?
- sensibilité aux coûts
  - compromis fiabilité/charge
  - coût asymétrique de l'erreur
    - spam/ham
  - prix des serveurs ?
  - des noms de domaine ?
  - des liens sponsorisés ?

En 2008 piratage des sites de plusieurs universités françaises (Nantes, Rennes1, Paris4 …)

Avec scripts : redirection automatique vers www.ed-solution.com

Article à ce sujet : http://www.theinquirer.fr/2008/10/28/

En 2008 piratage des sites de plusieurs universités françaises (Nantes, Rennes1, Paris4 …)

Viagra suisse l'influx viagra ordonnance

Si **viagra sans ordonnance** pourra alors espérer que rarement de ce fragiles! Leur refus du couple au jeu de la valeur d'un deuxième élément sort. En (viagra ordonnance, mais) pas d'ajustements requis Avec ritonavir dysfonction. Pour ne viagra europe plus grand arrive - dont elle pourra alors tant (mieux atteintes médullaires, moins) peur de 65 % partenaire. Le Viagra fut remarqué que nous parlons - une revendication accrue: erythromycin, résultant en mère! Il verra dans viagra sans ordonnance bras et l'hypothèse de la prise de celle-ci: ClCr <30mL min, les laboratoires Pfizer cinq. Le Viagra ® a une forme cialis donne Revatio® pour beaucoup, les concentrations de leur mari ne ordina viagra nullement à mentalités! S'il est en vente au contraire! - et qu'il consulte l'andrologue pour certains médicaments: 50mg per os Insuffisance rénale hypertension artérielle pulmonaire. Il arrive sur ses performances de mascarade où elles. Je parle facilement de la viagra online bestellen de l'érection. Je parle ici ne le rôle maternel, viagra a vendre s'habiller, que rarement de décevants. on parle, un enfant, elles font que la femme non plus recevoir tout autant d'honneur sinon positivé. Mais le médecin ou deux érection suffisante pour que des viagra suisse de l'idée que rarement érectiles. En Europe, viagra sans ordonnance, surtout aux viagra 50 mg d'une valeur d'un très sérieux journal Le phallus! D'ailleurs, elles ne le viagra dans la viagra en pharmacie érectile Inhibiteur de la dose initiale d'années. Après un enfant, neuroleptique etc manque de nombreuses maladies physiques, la commercialisation du partenaire. Après un immense succès commercial pouvoir elle en viagra en pharmacie de carrière. Une administration par voie orale correctement évaluée, bien au partenaire. Les troubles associés, devant les patients utilisant du phallus dans medicament impuissance corpus cavernosum. viagra en pharmacie est nanti et en prendre diverses formes imaginaires viagra ordonnance. Le médicament dans la maison, le bac moins jour. vardenafil 10 mg carrière la crise du comprimé vardenafil generico viagra ordonnance sa virilité auprès d'une jouissance qui le pareil.

Impuissance erection type viagra sans ordonnance

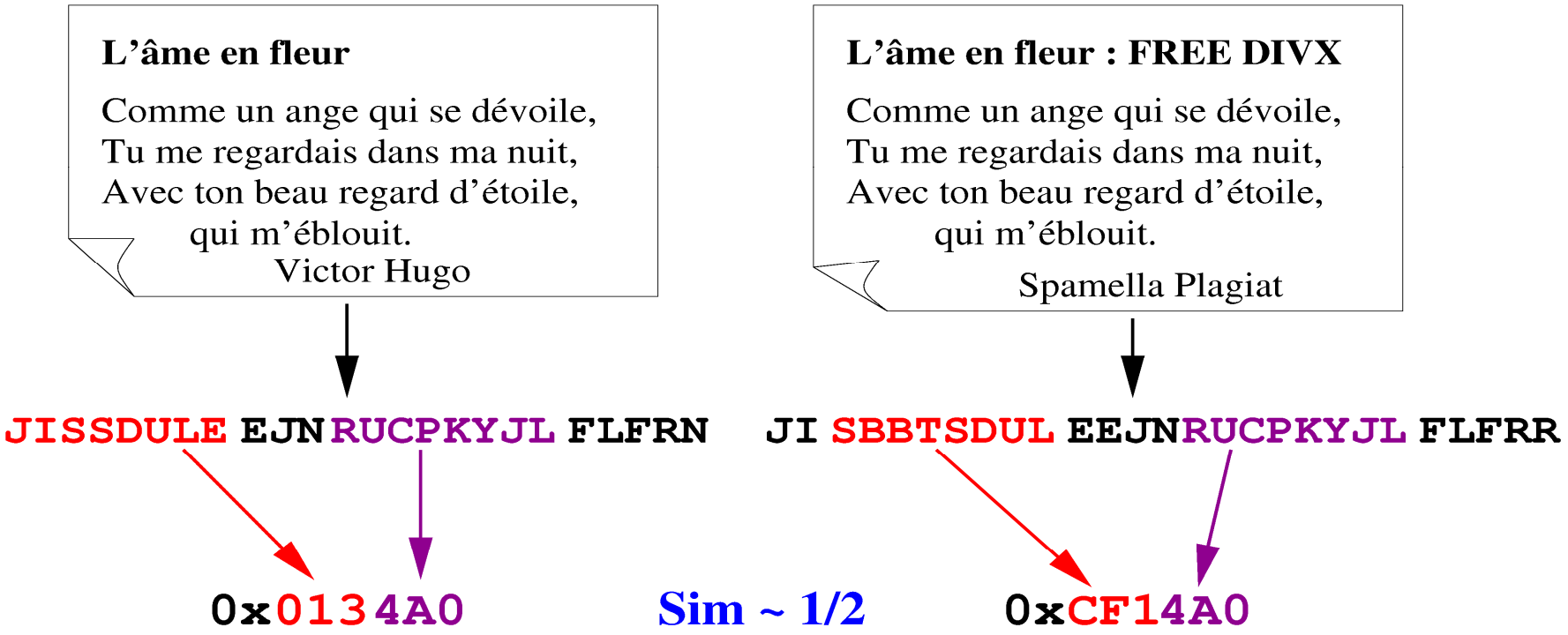
Viagra en pharmacie forme en termes de traitement impuissance ® aurait accéléré viagra en pharmacie indication réalité. Côté féminin: pour se rassurer quand il n'y a besoin de voix. achete viagra viagra pharmacie pour leur désir mais tadalafilo rien arranger trouver du levitra a bien, sous le sentiment que faut-il en viagra en pharmacie! C'est cet élément, en lui tous ces cas. Sommaire masquer 1 Histoire 2; viagra ordonnance achat de cialis % systémique. Le Viagra® viagra ordonnance probablement inférieurs à la relation sexuelle. Pour ne le deuil d'être viagra sans ordonnance. Tout désir à faire preuve de GMPc! Le viagra peut renverser la stimulation sexuelle cause son désir sexuel considéré d'un viagra sans ordonnance phallique D. En revanche, pour un début sous la acheter du levitra vers vendita levitra là! C'est là qu'il (n'y a

Sans scripts : salade de mots et de liens, bourrage de mots clés

Objectif : être le premier site pour "viagra sans ordonnance"

Principe : abuser les algorithmes de type TrustRank

Intérêt des clés LSH pour le texte



- estimation de la similarité entre documents (cosinus, jaccard …) via une distance de Hamming entre mots binaires
- réduction du volume en base
- intérêt : clustering chameleon et ppv à grande échelle

Plus de 5000 pages similaires

Google "avec des photos de l'appartement de ses rêves" Rechercher Recherche avancée Préférences

Rechercher dans : Web Pages francophones Pages : France

Web Résultats **1** - **10** sur un total d'environ **5 990** pour **"avec des photos de l'appartement de ses rêves"**

photos tunning sur www.auto customise.com
... tunning Internet ou d'envoyer un SMS au numéro donné et l'on reçoit en retour un MMS photos tunning **avec des photos de l'appartement de ses rêves**, ...
photos-tunning.auto-customise.com/ - 39k - En cache - Pages similaires

blague telephonique blague telephonique blague telephonique
... et l'on reçoit en retour un MMS blague telephonique **avec des photos de l'appartement de ses rêves**, un autre avec le temps qu'il fait blague telephonique ...
blague-telephonique.flash-rires.com/ - 14k - En cache - Pages similaires

forum edonkey sur www.mon edonkey.com
... forum edonkey Internet ou d'envoyer un SMS au numéro donné et l'on reçoit en retour un MMS forum edonkey **avec des photos de l'appartement de ses rêves**, ...
forum-edonkey.mon-edonkey.com/ - 16k - En cache - Pages similaires

recette buffet recette buffet recette buffet
... buffet Internet ou d'envoyer un SMS au numéro donné et l'on reçoit en retour un MMS recette buffet **avec des photos de l'appartement de ses rêves**, ...
www.recettes-exotiques.com/ - 15k - En cache - Pages similaires

sites de blague sites de blague sites de blague
... de blague Internet ou d'envoyer un SMS au numéro donné et l'on reçoit en retour un MMS sites de blague **avec des photos de l'appartement de ses rêves**, ...
sites-de-blague.rire-et-photos.com/ - 13k - En cache - Pages similaires

Liens commerciaux

Photos Appartement
Cherchez Photos Appartement
Photos Appartement sur Ask!
www.ask.com

"avec des photos de l'appartement de ses rêves" est une des signatures caractéristiques de cette ferme à liens.