

Méthodes automatiques pour la détection du spamdexing

Orange Labs

Tanguy Urvoy, Recherche & Développement
26 mars 2009, DAPA



diffusion libre





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)



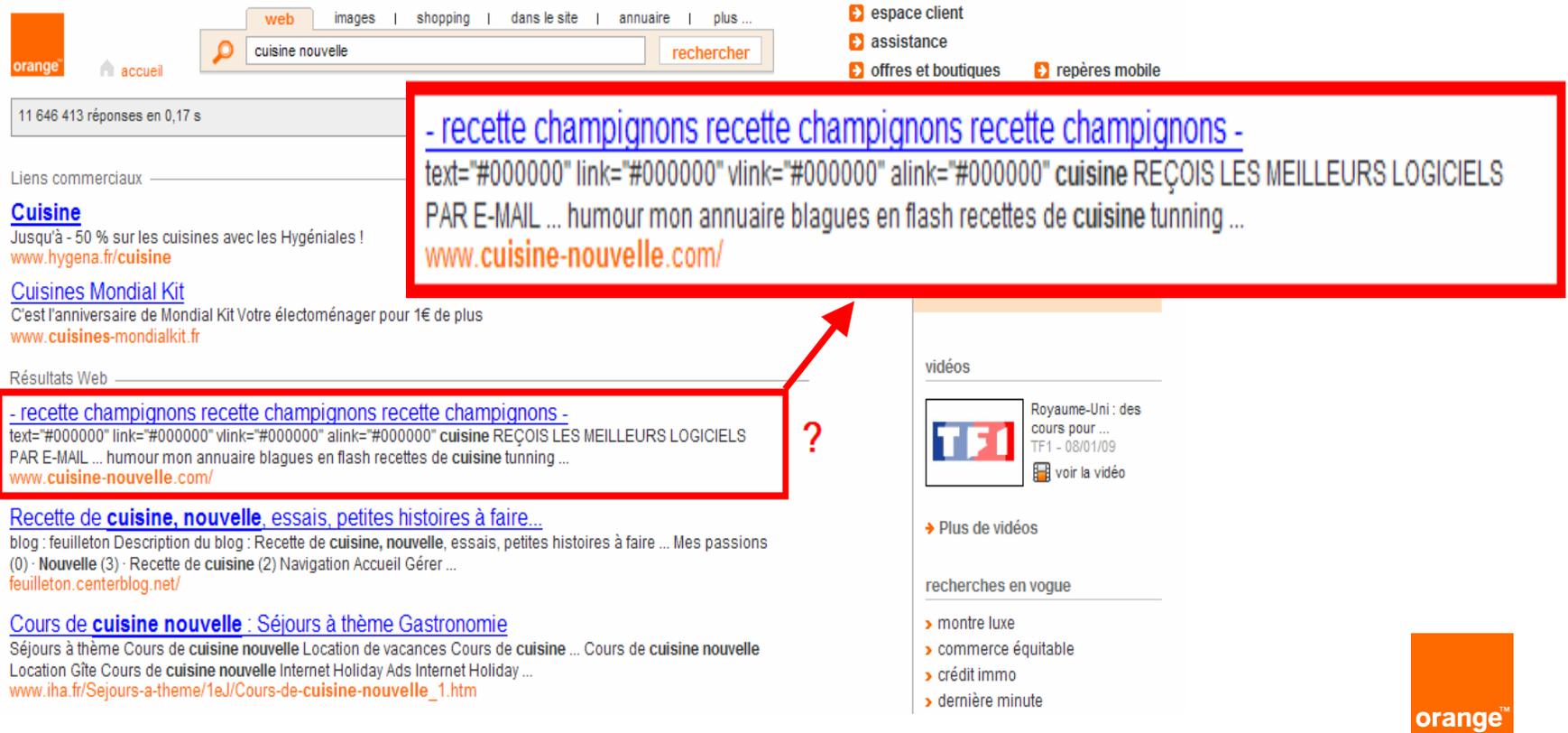


Spamdexing



Définition

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



orange | accueil | recherche | cuisine nouvelle | rechercher

espace client | assistance | offres et boutiques | repères mobile

11 646 413 réponses en 0,17 s

Liens commerciaux

[Cuisine](#)
Jusqu'à - 50 % sur les cuisines avec les Hygénéales !
www.hygena.fr/cuisine

[Cuisines Mondial Kit](#)
C'est l'anniversaire de Mondial Kit Votre électroménager pour 1€ de plus
www.cuisines-mondialkit.fr

Résultats Web

[- recette champignons recette champignons recette champignons -](#)
text="#000000" link="#000000" vlink="#000000" alink="#000000" cuisine REÇOIS LES MEILLEURS LOGICIELS
PAR E-MAIL ... humour mon annuaire blagues en flash recettes de cuisine tuning ...
www.cuisine-nouvelle.com/

[Recette de cuisine, nouvelle, essais, petites histoires à faire...](#)
blog : feuilleton Description du blog : Recette de cuisine, nouvelle, essais, petites histoires à faire ... Mes passions
(0) · Nouvelle (3) · Recette de cuisine (2) Navigation Accueil Gérer ...
feuilleton.centerblog.net/

[Cours de cuisine nouvelle : Séjours à thème Gastronomie](#)
Séjours à thème Cours de cuisine nouvelle Location de vacances Cours de cuisine ... Cours de cuisine nouvelle
Location Gîte Cours de cuisine nouvelle Internet Holiday Ads Internet Holiday ...
www.ihh.fr/Séjours-a-theme/1eJ/Cours-de-cuisine-nouvelle_1.htm

vidéos

 Royaume-Uni : des cours pour ...
TF1 - 08/01/09
[voir la vidéo](#)

Plus de vidéos

recherches en vogue

- montre luxe
- commerce équitable
- crédit immo
- dernière minute



Ferme à liens, cuisine et illusions



DIXMILLErecettes

menu

- apéritifs (584)
- accompagnements (785)
- plats principaux (1130)
- soupes (631)
- végétariennes (1075)
- pâtes (905)
- poulet (644)
- crustacés (454)
- viandes (512)
- porc (548)
- petits déjeuners rapides et faciles (354)
- desserts (788)
- gâteaux (454)
- boissons et cocktails (624)

RECETTES DISPONIBLES: 10.000

références
 (V) = VÉGÉTARIENNES (S) = SPÉCIAL (O) = FROIDES (H) = CHAUDES (CH) = CHINE

Bienvenue sur le meilleur site de recettes du Web ! Qu'est-ce que vous voulez préparer aujourd'hui ?

nous recommandons... des plats spéciaux pour le plaisir de tous.

- Salades**
Etonnez votre famille avec ces délicieuses salades
- Betteraves et pommes de terre (v)
- Salade d'anchois
- Salade Impérial (v)
[en savoir plus...](#)
- Pâtes et Sauces**
Les traditionnelles pâtes italiennes avec des sauces délicieuses
- Pâtes aux noix (v)
- Pâtes frites
- Spaghetti
[en savoir plus...](#)
- Gâteaux**
Délicieux gâteaux pour les plus gourmands.
- Muffins aux prunes
- Gâteau noir (fruit)
- Gâteau à la cannelle et aux carottes
[en savoir plus...](#)
- Desserts**
Froids et chauds, glaces, desserts aux fruits, etc...
- Gâteau glacé au chocolat et au rhum
- Mousse au chocolat
- Rougat glacé (v)
[en savoir plus...](#)
- Gâteaux aux fruits**
Facile et rapide, ce délicieux gâteau aux fruits va régaler vos invités.
[La recette ici](#)
- Soupe printemps**
Savoureuse, faible en calories et très nutritive cette soupe de légumes est idéale pour le dîner.
[La recette ici](#)

apéritifs | plats principaux | accompagnements | desserts | boissons | rapides et faciles

Entrée Membres | Politique de Confidentialité | Conditions d'Utilisation | Webmasters | Service Client

Copyright © 2002 - 2006 Live Interactive S.A. Tous Droits Réservés.



TOUTES LES ILLUSIONS

RECHERCHE DES EFFETS D'OPTIQUE

Bienvenue sur le site des effets d'optique les plus incroyables. Etonne tes copains et perds le sens des réalités. Des illusions les plus trompeuses !

- 3 SUR 1**
Découvre ces images au triple sens. Tout dépend de quel côté tu regardes.
- CAMOUFLAGE**
Il y a un secret occulté dans ces images. Les vois-tu ?
- FOND ET FORME**
Essaie de trouver les limites entre le fond et la forme ? Attention, ce n'est pas toujours facile...
- IMPOSSIBLES**
Tu vas te casser la tête avec ces images ! As-tu le courage d'essayer ?
- SPIRALE INFERNALE**
Attention ! Dangereux pour les cardiaques ! A éviter si tu as le vertige !
- GÉOMETRIE**
Toutes les règles ont des exceptions ! Quelques unes de ces lignes pourraient te tromper.
- ARCHITECTURE**
Figures et structures qui défient toutes les lois de la gravité.
- VISAGES**
Tous les secrets qui peuvent se cacher derrière les visages...
- LETTRES**
Découvre ce qui est caché derrière les lettres de l'alphabet ! Tu vas être étonné.

ENTRER ! PLUS D'ILLUSIONS

Entrée Membres | Politique de Confidentialité | Conditions d'Utilisation | Webmasters | Service Client

Copyright © 2002 - 2006 Live Interactive S.A. Tous Droits Réservés.

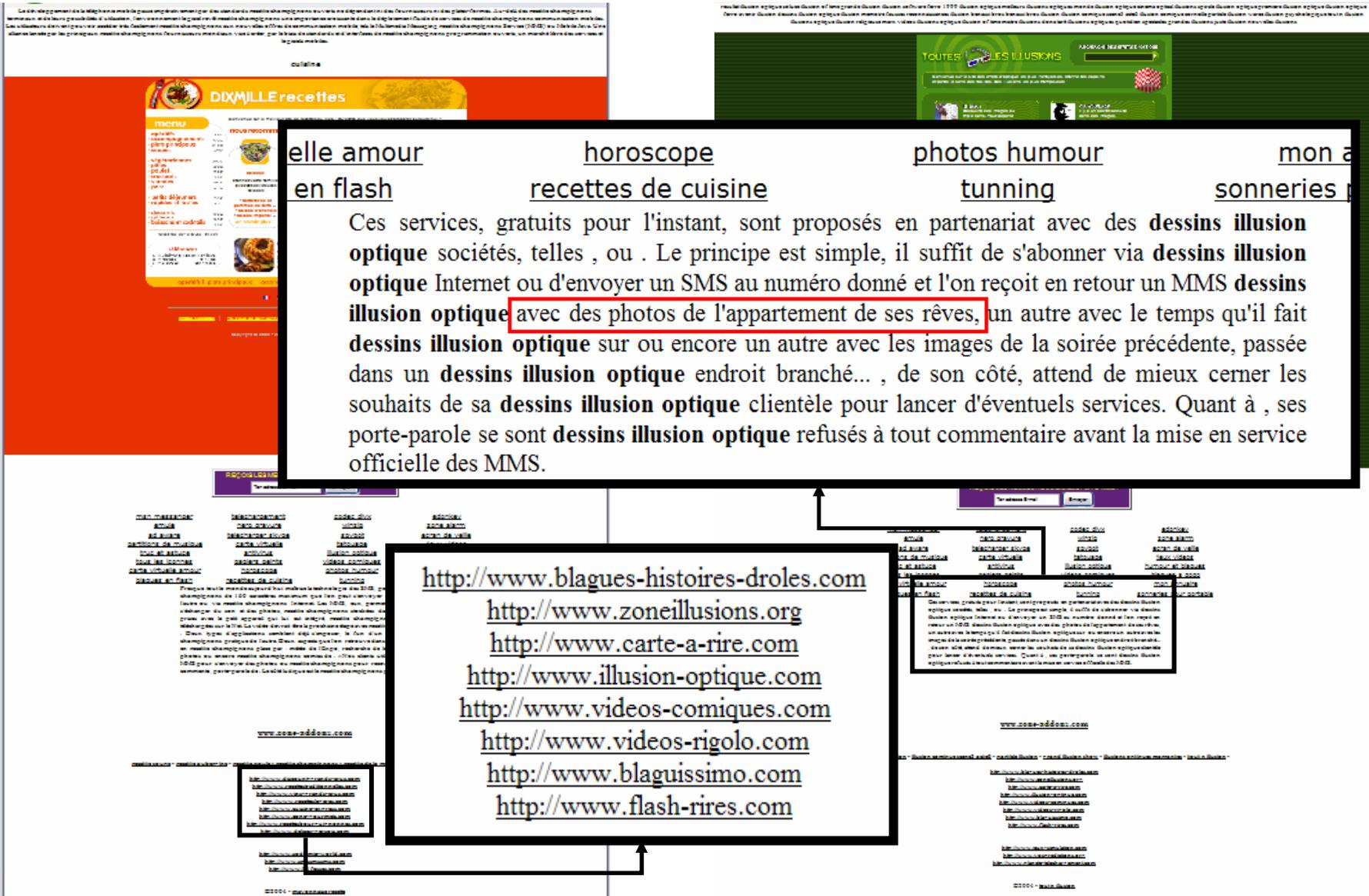
- le site qui parle (entre autre) de champignons : cuisine-nouvelle.com
- fortement connecté avec des sites similaires comme mes-illusions.com
- ces sites n'offrent pas de contenu réel :



- ils sont générés de manière automatique
- c'est une **ferme à liens** ventilée sur plusieurs milliers de domaines.



Camouflage javascript, 8-gramme caractéristique



elle amour en flash horoscope recettes de cuisine photos humour tuning mon a sonneries

Ces services, gratuits pour l'instant, sont proposés en partenariat avec des **dessins illusion optique** sociétés, telles , ou . Le principe est simple, il suffit de s'abonner via **dessins illusion optique** Internet ou d'envoyer un SMS au numéro donné et l'on reçoit en retour un MMS **dessins illusion optique** avec des photos de l'appartement de ses rêves, un autre avec le temps qu'il fait **dessins illusion optique** sur ou encore un autre avec les images de la soirée précédente, passée dans un **dessins illusion optique** endroit branché... , de son côté, attend de mieux cerner les souhaits de sa **dessins illusion optique** clientèle pour lancer d'éventuels services. Quant à , ses porte-parole se sont **dessins illusion optique** refusés à tout commentaire avant la mise en service officielle des MMS.

<http://www.blagues-histoires-droles.com>
<http://www.zoneillusions.org>
<http://www.carte-a-rire.com>
<http://www.illusion-optique.com>
<http://www.videos-comiques.com>
<http://www.videos-rigolo.com>
<http://www.blaguissimo.com>
<http://www.flash-rires.com>

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

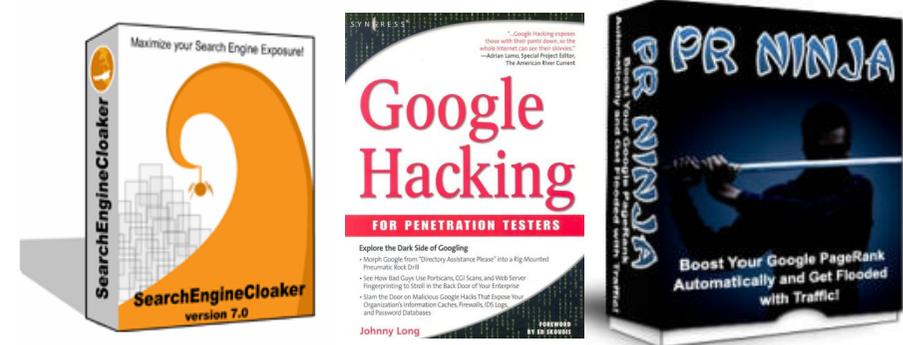
- gaspille de la bande passante lors du crawl
- pollue les bases et les index
- dégrade la pertinence des listes réponse





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



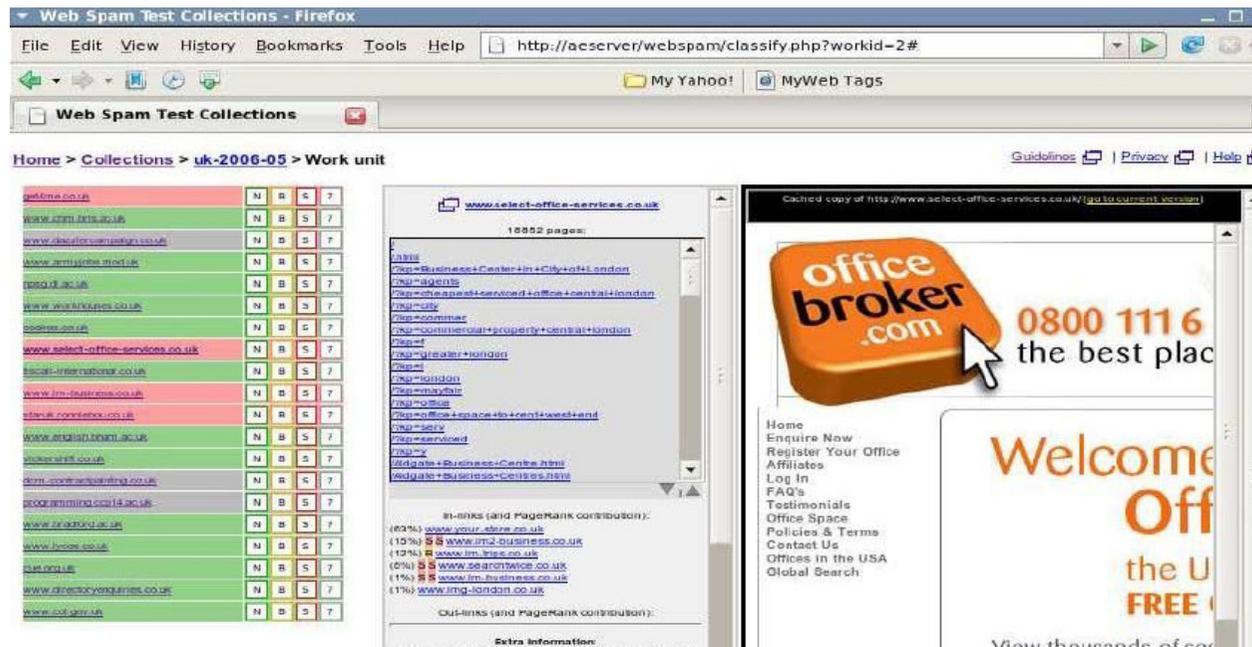


Remarques générales sur le spamdexing

- **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 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



The screenshot shows a Firefox browser window titled 'Web Spam Test Collections - Firefox'. The address bar contains 'http://aeserver/webspam/classify.php?workid=2#'. The page content includes a navigation menu with 'Home > Collections > uk-2006-05 > Work unit'. Below this is a table of URLs with columns for 'N', 'B', 'S', and '7'. The table lists various URLs, some of which are highlighted in red or green. To the right of the table is a preview of a webpage for 'www.select-office-services.co.uk', showing a list of 18852 pages and a search bar. The preview also shows a large orange button that says 'office broker.com' and a mouse cursor pointing to it. Below the button, there is text that says '0800 111 6' and 'the best plac'. The page also features a 'Welcome Off' banner and a 'FREE' offer.

- 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

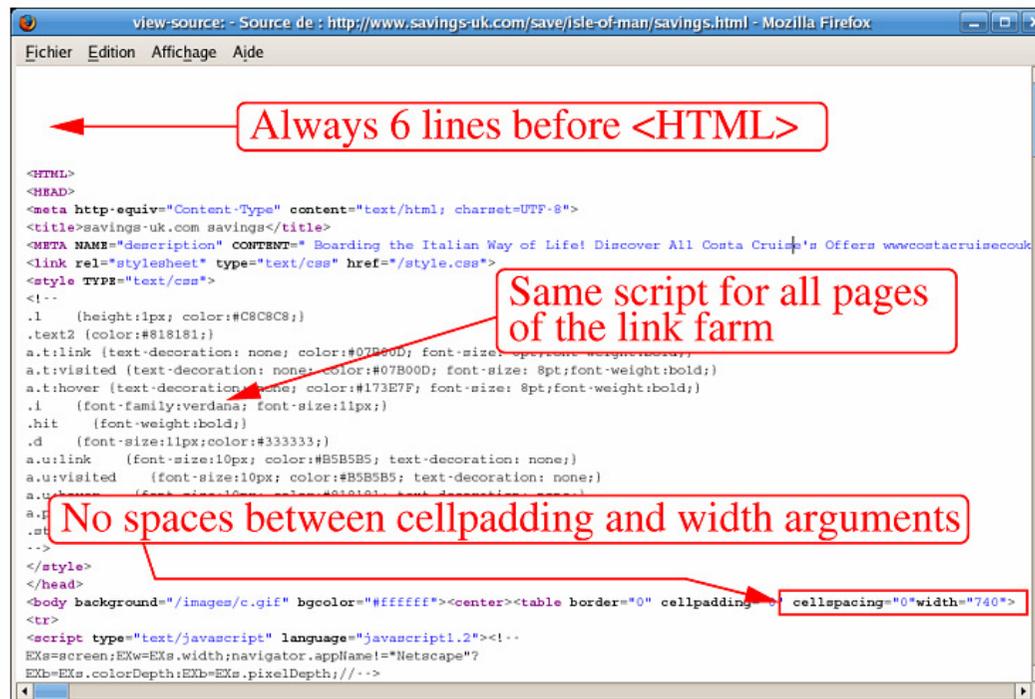
- les scripts perl manquent de créativité

get4me.co.uk (index)

www.goto4me.co.uk

www.injuryclaim-uk.com

4u-insurance.co.uk



```
<HTML>
<HEAD>
<meta http-equiv="Content-Type" content="text/html; charset=UTF-8">
<title>savings-uk.com savings</title>
<META NAME="description" CONTENT=" Boarding the Italian Way of Life! Discover All Costa Cruise's Offers www.costacruise.co.uk" >
<link rel="stylesheet" type="text/css" href="/style.css">
<style TYPE="text/css">
<!--
.l (height:1px; color:#C8C8C8;)
.text2 (color:#818181;)
a.t:link (text-decoration: none; color:#07B00D; font-size: 8pt; font-weight: bold;)
a.t:visited (text-decoration: none; color:#07B00D; font-size: 8pt; font-weight: bold;)
a.t:hover (text-decoration: none; color:#173E7F; font-size: 8pt; font-weight: bold;)
.i (font-family: verdana; font-size: 11px;)
.hit (font-weight: bold;)
.d (font-size: 11px; color: #333333;)
a.u:link (font-size: 10px; color: #B5B5B5; text-decoration: none;)
a.u:visited (font-size: 10px; color: #B5B5B5; text-decoration: none;)
a.u:hover (font-size: 10px; color: #818181; text-decoration: none;)
a.p (font-size: 10px; color: #818181; text-decoration: none;)
-->
</style>
</head>
<body background="/images/c.gif" bgcolor="#EEEEEE"><center><table border="0" cellpadding="0" cellspacing="0" width="740">
<tr>
<script type="text/javascript" language="javascript1.2"><!--
EXs=screen;EXw=EXs.width;navigator.appName!="Netscape"?
EXb=EXs.colorDepth;EXb=EXs.pixelDepth;!-->
```

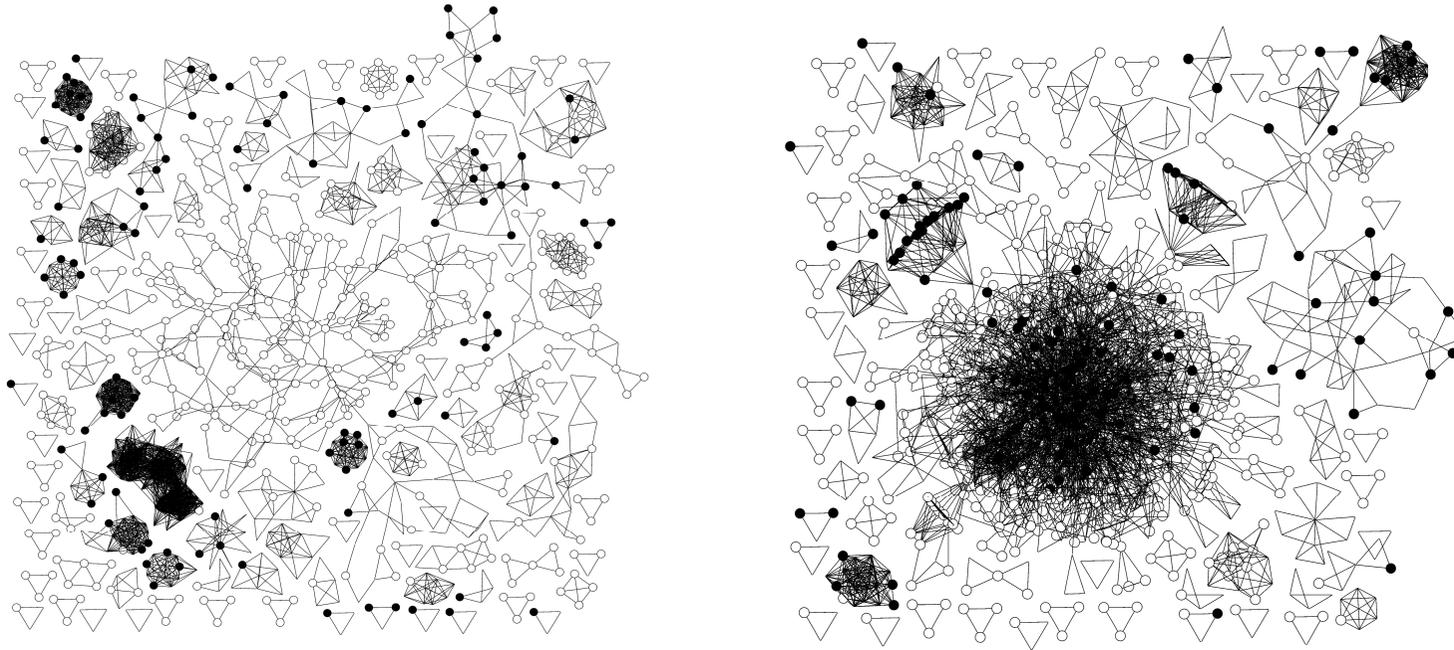
Always 6 lines before <HTML>

Same script for all pages of the link farm

No spaces between cellpadding and width arguments



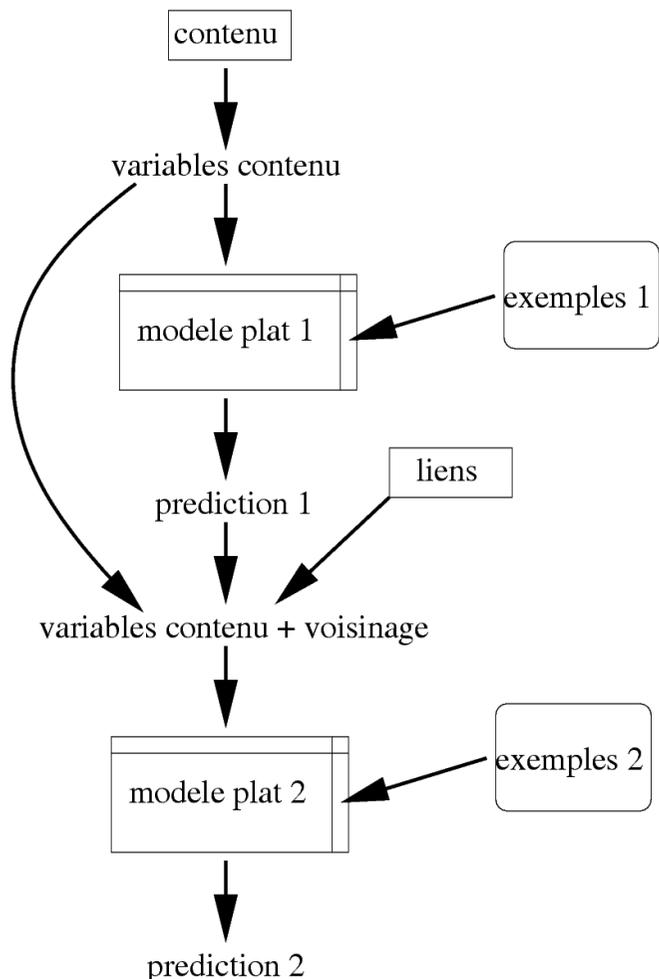
Propagation d'étiquettes : analyse liens/contenus



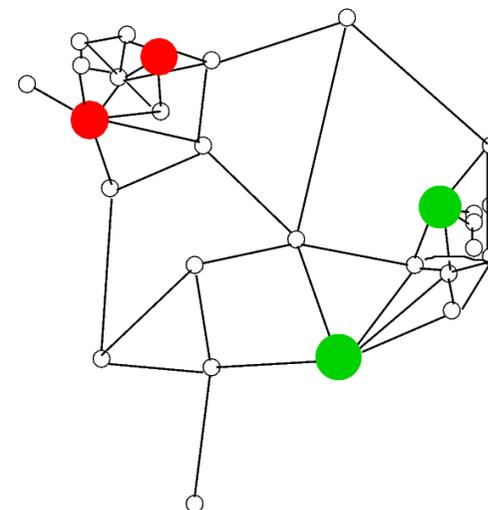
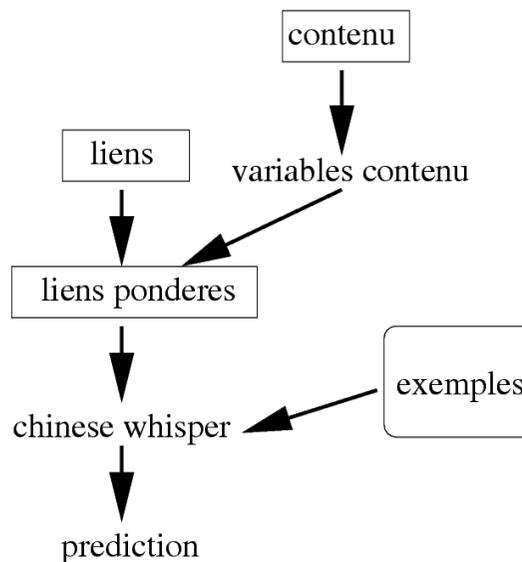
- graphes partiellement étiquetés + contenus
- nombreuses méthodes
 - renforcement du diagnostic contenu par voisinage
 - pondération des arcs selon contenu et propagation
- workshop graphlab 2007 et challenges WebSpam



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

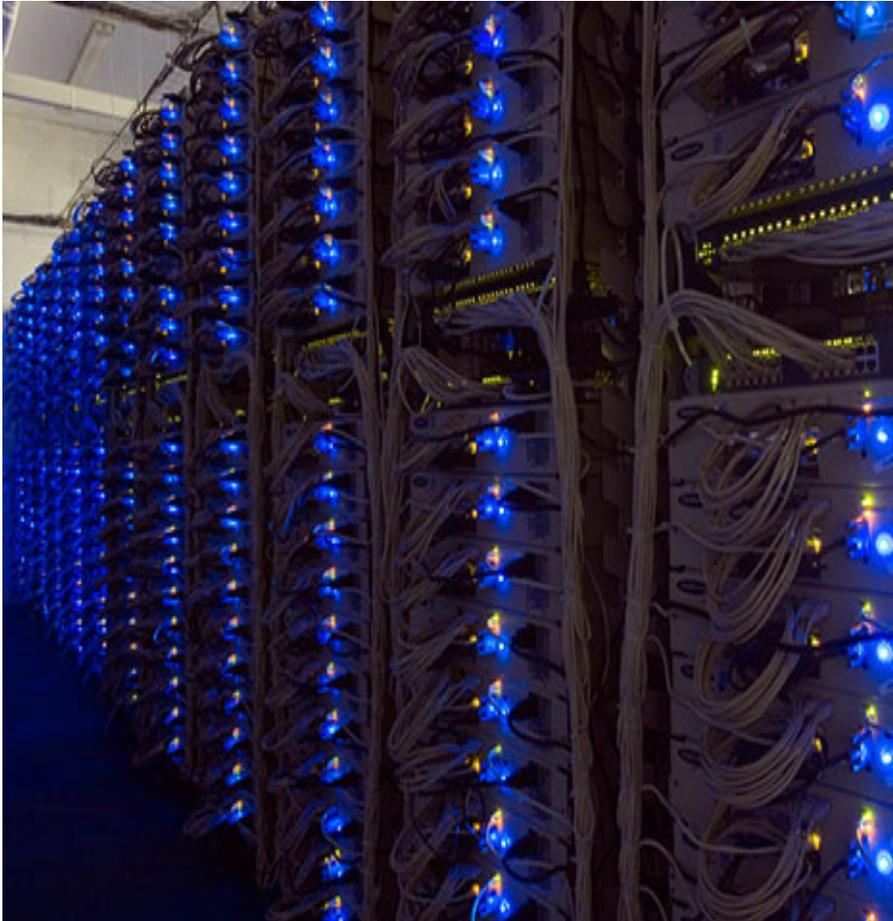




Annexes



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 ...)



The screenshot shows a website for 'VIAGRA ONLINE STORE'. At the top, there's a navigation bar with links: ACCUEIL / PRODUITS, QUESTIONS FRÉQUENTES, NOTRE POLITIQUE, and CONTACTEZ-NOUS. The main banner features a couple embracing and a white dove, with the word 'SANTÉ' prominently displayed. To the right, a promotional box offers 'Quelle affaire! €75.01' for 'ViagraSoft + CialisSoft' (10 pills each). Below this, a list of benefits includes: 'PAS DE VISITES MÉDICALES', 'LES PLUS BAS PRIX DANS LE MONDE ENTIER', 'EMBALLAGE DISCRET', 'AUCUN EMBARRAS', 'ESCOMPTES HEBDOMADAIRES', and 'ACHETEZ DANS LE BULD ET ÉCONOMISEZ'. The 'Produits' section lists 'VIAGRA' (Sildenafil Citrate 100mg) for €1.01 per pill and 'CIALIS' (Tadalafil 20mg) for €1.27 per pill, each with an 'ACHETEZ' button. A 'Témoignages de nos clients' section contains a testimonial about a customer's experience with the service.

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 ritoonavir 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 positif. 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 [médicament 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 [tadalafil](#) 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

L'âme en fleur
 Comme un ange qui se dévoile,
 Tu me regardais dans ma nuit,
 Avec ton beau regard d'étoile,
 qui m'éblouit.
 Victor Hugo

L'âme en fleur : FREE DIVX
 Comme un ange qui se dévoile,
 Tu me regardais dans ma nuit,
 Avec ton beau regard d'étoile,
 qui m'éblouit.
 Spamella Plagiat

JISSDULE EJNRUCPKYJL FLFRN

JI SBBTSDUL EEJNRUCPKYJL FLFRR

0x0134A0

Sim ~ 1/2

0xCF14A0

- 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



"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 tuning sur www.auto customise.com](#)

... tuning Internet ou d'envoyer un SMS au numéro donné et l'on reçoit en retour un MMS photos tuning **avec des photos de l'appartement de ses rêves**, ...

[photos-tuning.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.

