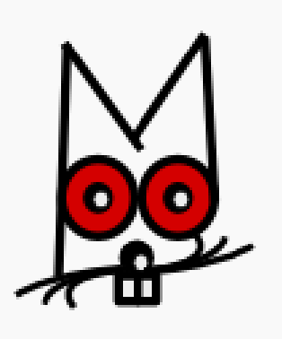


Google est multidimensionnel

Google est multidimensionnel

Cette page fait partie de l'explication sur le fonctionnement de google

Vous avez certainement entendu parler de l'indice de confiance d'un site. Il semblerait même que ce soit plus important que le page rank d'après les pseudos experts en référencement qu'il est facile de rencontrer (ou qui se jettent à vos genoux) dès qu'on a un site pourri et beaucoup d'argent à dépenser. De mon point de vue, le page rank publié est plus que déprécié, et l'indice de confiance n'est qu'une résultante de la façon dont google fonctionne actuellement.

Google est un hypercube

Google possède un ensemble de dimensions indépendantes dans lesquelles il place chaque page indexée. Prenons un exemple simple en 3 dimensions, ce qui va me permettre de faire des schémas assez moches sans trop de problème, mais le principe reste le même en dimension 16 ou 38. Je m'intéresse à la requête « sport » pour l'illustration.

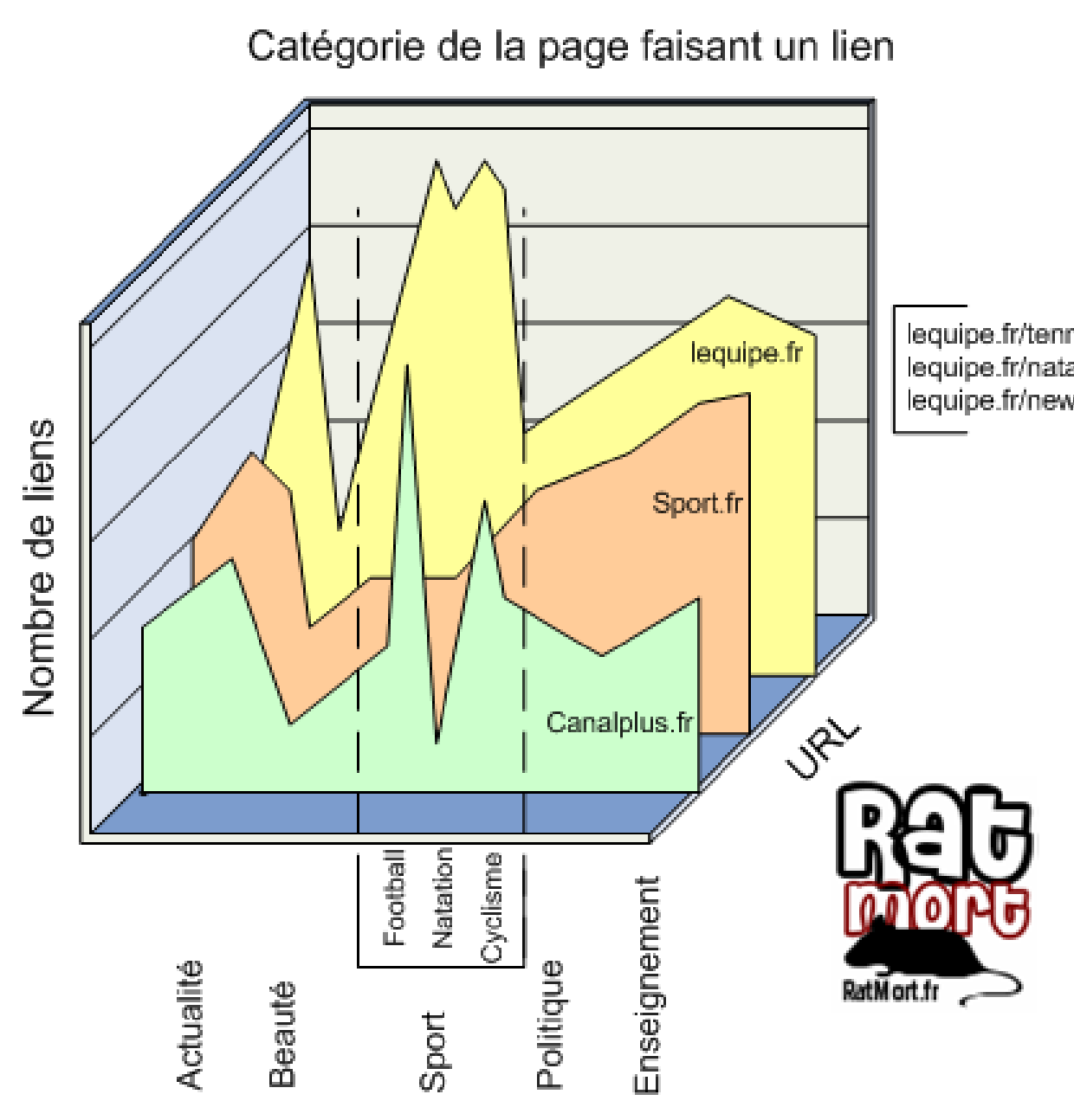
Google en 3D

Considérons google comme une structure vectorielle de dimension N que l'on fixe arbitrairement à 3 ici. La première dimension est par exemple le domaine ou secteur de la page qui effectue un lien retour (backlink). Il s'agit de la catégorie, comme proposée dans google ad planner.

Pour le site lequipe.fr, les informations fournies par google planner sont les suivantes :

```
Sports
News & Current Events > Magazines
Lifestyles > Men's Interests
```

Google communique au sujet des catégories par site mais conserve des catégories par pages au moins sur les pages à plus fort trafic (celles qui apparaissent dans les site-links par exemple). Attention, les propriétaires de site peuvent modifier ce qui s'affiche dans google ad planner (mais pas la catégorie que google utilise dans le classement des sites). Prenons comme deuxième dimension l'url de la page.



Chaque catégorie n'est pas uniquement une liste, mais une hiérarchie organisée (proche d'un modèle orienté objet, un annuaire ldap ou un fichier xml) que l'on appelle ici arbre. Google possède aussi des dictionnaires de langue avec les mots classés de la même façon. Par exemple, dans animaux, il y a mammifère et ovipare, et dans mammifère, on retrouve bipède, quadrupède Chaque langue est un arbre, et un autre type d'arbre (appelons le arbre de traduction) fait correspondre partiellement les mots d'une langue à une autre. En interne, le term utilisé est "umbrella", c'est-à-dire parapluie, et google maniplue par exemple un "umbrella of synonyms".

Dans notre exemple, seule la catégorie sport a été détaillée ainsi que les URLs sur le site lequipe.fr, afin d'expliquer le mécanisme. Cet arbre est très probablement généré en partie grâce aux requêtes des internautes, car il comprend des mots comme « horaire cinéma » ou « demande de stage » qui ne sont pas à proprement parlé des composants d'un dictionnaire officiel. Cet arbre est aussi utilisé pour proposer des requêtes proches en pied de page, ou aider à corriger les fautes de frappe lors d'une recherche.

Considérons de plus comme troisième dimension, le nombre de liens vers la page. Google utilise plutôt le nombre de liens provenant de sites de la catégorie sport et ayant une ancre dont le mot est situé dans une sous-branche de la branche sport du dictionnaire français, mais nous pouvons considérer comme troisième dimension le nombre de liens retour total comme première approximation.

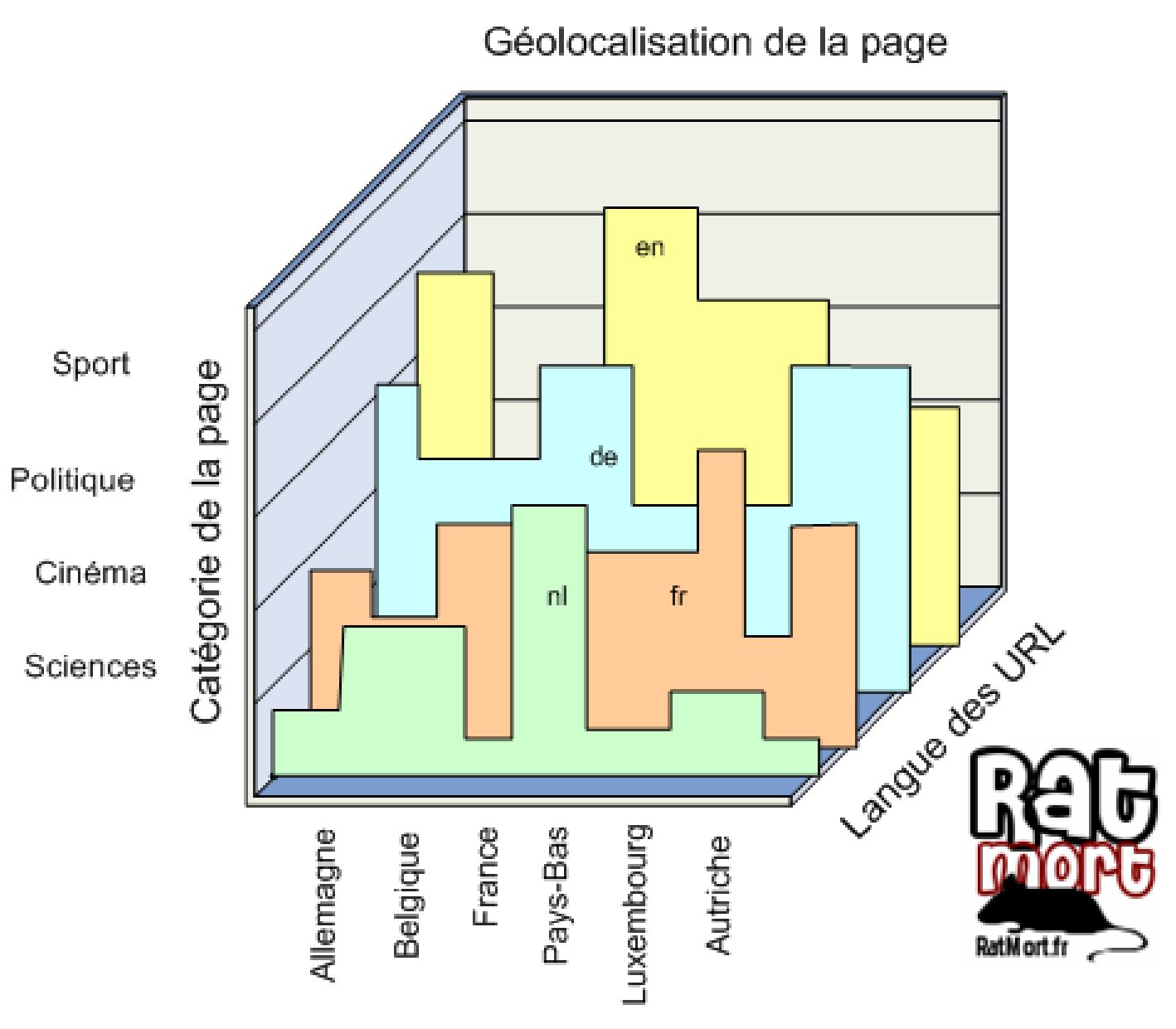
Chaque site est positionné dans ce cube selon les 3 critères définis. Si vous cherchez « sport » dans google, le cube est projeté selon les axes concernés. On obtient ainsi une liste d'urls classées en fonction des différents plans considérés. C'est sur base de cette liste que google compose sa page de résultat.

La question que vous devriez vous poser à ce moment de l'exposé est la suivante: d'où viennent les données du 3ème axe, le nombre de liens retour, par catégorie? Et bien elles proviennent d'une autre projection dans un autre cube, qui liste tous les backlinks sur l'axe z, la catégorie sur l'axe y et l'ancre sur l'axe x. L'axe x n'est pas linéaire (c'est-à-dire n'est pas une simple liste) mais un arbre. De même, la liste des backlinks est composée d'un arbre décrivant les différentes URLs. La projection peut se situer à n'importe quel niveau de l'arbre, soit à la racine du site, soit au niveau d'une page ou d'un composant de page (une image par exemple, même si les résultats sont moins probant pour le moment. En effet, considérer une image comme un sous-ensemble d'une page associe trop fortement le texte entourant l'image ou le thème de la page au contenu graphique de l'image).

On pourrait considérer cette structure comme un cube dans lequel le niveau de zoom n'a aucune incidence sur les fonctionnalités d'analyse de l'information qu'il propose. Google est multi-niveaux et multi-dimension. Il s'agit d'un espace vectoriel de dimension n avec des caractéristiques fractales.

Google et la géolocalisation

Je vous propose un autre exemple qui utilise 2 autres dimensions pour lesquelles google est très bon : la détection de langue et la géolocalisation. L'illustration suivante propose un cube conçu pour une requête du genre « film avatar lyon ». Ce cube permet en effet d'effectuer une projection selon les bons axes et de proposer une droite vectorielle qui est un fait une liste de pages en français, proche du domaine du cinéma, et ayant un rapport avec la région lyonnaise.



Il est important de rappeler ici que l'axe de géolocalisation est un arbre, et qu'il est donc hiérarchique (continents, pays, régions, villes.

Fonctionnement de Google Map

Petite remarque à l'attention des utilisteurs de google map. Si vous cherchez une petite rue près de chez vos grand-parents à 800 kms de chez vous, mais sans fournir de code postal ni de ville dans google map, vous tomberez certainement sur une rue portant le même nom d'une grande ville plus ou moins proche de chez vous, ou aux Etats-Unis si google ma pets vraiment largué. Si maintenant vous cherchez la rue principale du village de vos grands-parents, et effectuez ensuite un calcul d'itinéraire vers la petite rue sus-mentionnée, toujours sans fournir plus d'information que le nom de la rue et le numéro, google map trouve sans problème. La deuxième fois, google part de la bonne branche et cherche aux alentours, jusqu'à trouver la rue, en remontant dans l'arbre de géolocalisation. La requête est identiquement construite pour tous les utilisateurs, mais le résultat diffère selon la branche dans laquelle google nous situe.

Pour le moteur de recherche, c'est la même chose. Si vous êtes francophone, votre recherche part de la branche francophone.

Dans le futur, si vous cherchez régulièrement des informations sur du Perl et du C++, lorsque vous chercherez Python ou Java, vous aurez moins de résultats sur les serpents et la danse que vos voisins, sauf si vos voisins travaillent chez IBM

Les axes utilisés par google

D'autres axes peuvent être considérés et sont en train d'être mis en uvre par google, comme par exemple le champ lexical et le type de vocabulaire (ordurier, grossier, familier, courant, soutenu, littéraire) qui est un très bon moyen de démasquer facilement les sites non destinés aux enfants. Je pense que vous pouvez par vous-même imaginer les axes que google utilise actuellement. Certains sont publiés par google, d'autres sont sous-entendus:

- type de document : mp3, html, pdf, avi
- type de site : forum, blog, application en ligne, site classique
- encodage utilisé : UTF-8, ... a priori n'a pas d'impact sur le positionnement
- vitesse de réponse du serveur, pris en compte partiellement pour le moment

Certaines dimensions n'ont pas d'impact sur le classement dans les résultats de recherche.

Indice de confiance de google : indice de hilltop

Une dimension particulière qu'il est intéressant de présenter ici est en fait une dimension de second niveau, c'est-à-dire qu'elle peut être déduite des dimensions indépendantes qui forment le cube (c'est une combinaison linéaire d'autres dimensions de base). Il s'agit de l'indice de confiance que j'appellerais **indice de Hilltop**, calculable par un algorithme de HITS. En se basant sur les catégories des sites pointant vers une page donnée et de leur propre indice de hilltop, on en déduit l'indice de hilltop de la page analysée. Cette méthode récurrente permet d'assigner un indice de hilltop à toutes les pages indexées qui servira à pondérer les résultats récupérés grâce à une projection dans le cube.

Cet indice est directement corrélé avec la catégorie, ce qui n'était pas le cas avec la pagerank historique bien connu. En d'autres termes, il existe un indice de hilltop pour chaque catégorie (sport, actualité, ...).

Remarque sur l'indice de hilltop à l'attention des référenceurs

Cet indice étant lié à une catégorie, un même site avec un même contenu peut être positionné différement en fonction de la catégorie assignée par google. Imaginons par exemple un site qui explique par les mathématiques quelles sont les probabilités de gain à la roulette. Si google considère que sur le mot roulette, la catégorie de cette page est "mathématiques", ou bien si au contraire il déduit du contenu que la catégorie la plus appropriée est "jeux d'argent", le positionnement peut être complètement différent.

La construction multidimentionelle de google fait qu'il considère que les sites parlant de mathématiques sont plus en général des sites de référence que ceux au sujet de jeux d'argents. Personne chez google n'a définit explicitement cette rêgle, mais elle est la conséquence de la façon dont google est conçu. Certains référenceurs utilisent le terme de voisinage, je parlerais plutôt de probabilité plus importante de crédibilité pour les sites parlant de mathématiques par rapport à ceux de casino, basée sur le contenu courant du cube.

Ce principe explique les "blacklistages" que l'on peut constater. En effet, un site qui propose des extraits de bandes-annoces de cinéma par exemple, est classé dans les catégories cinéma (80%), actualité (10%), streaming (10%). Le positionnement est bon.

Plus tard, quelques sites de streaming ajoutent des liens vers des bandes-annonces, et le webmaster ajoute explicitement en dessous de toutes ces pages "regardez la bande annonce de xxx en streaming". Cela induit un recalcul des catégories, qui est maintenant de 55% pour le streaming, 30 % cinéma et 10% actualité et 5% culture. Le site, après projection dans le cube, est maintenant défini comme un site de streaming sur l'axe des catégories. Il est très mal placé sur des requêtes comme "bandes annonces festival de cannes" ou "actualité cinéma vidéo". Par ailleurs, la requête "streaming" est squattée par des sites récupérant au moins des centaines de liens par semaine. Le jour où les petits sites faisant des liens vers ces gros sites ferment ou baissent le nombre de liens créés par unité de temps (quelques mois à quelques années plus tard, ou bien parfois jamais), le site original peut ressortir correctement sur "streaming" si jamais il a continué à recevoir des liens réguliers, mais il sera difficilement positionnable sur "cinéma".

Cet exemple rapide (et certainement mal choisi) explique une grande partie des blacklistages soit-disant aveugles de google et est l'origine d'une multitude de tentatives d'explications ~~bidons sur des forums de référenceurs ignares~~ non objectives.

Suite : Google et la gestion du temps réel

COMMENTAIRES

Si vous avez appris des trucs, je vous remercie de cliquer sur le bouton google plus

LISEZ CES ARTICLES!

- Comment être premier dans Google?
- Pourquoi les résultats de Google sont différents?
- Améliorer la vitesse de chargement du site
- Recette pour référencer son site internet
- White Hat vs Black Hat
- Optimiser ses landing pages
- Script php pour les échanges de liens
- Google - frequently mentioned on the web

Techniques Black Hat

- Adsense Clickjacking
- Cloaking indetectable pour tromper adsense
- Cloaking temporaire et erreur 404
- Bibliographie

Algorithme Google

1. Le fonctionnement de google
2. Google est multidimensionnel
3. Google et le temps reel
4. Google et son algorithme sont simples

QUI SUIS-JE?

Je suis Rat Mort. Je vais régulièrement dans les **locaux de Google** pour aller manger des donuts avec mes amis Pandas. Par ailleurs, je déploie et optimise des sites internet pour générer un traffic énorme... Et ça marche!

QUELQUES LIENS

Nous contacter
Offres d'emploi

CONTACT

Rat Mort
21 Av. des Arts,
1000, Bruxelles,
Belgium.
☆ Rat Mort ☆

ZONE MEMBRES

La section privée s'adresse aux référenceurs avertis

Email *

S'inscrire