Chronique

de la recherche sur le web

par Marc Duval 3 avril 2003, vol. 3, n° 4

La pertinence de Google : étude de cas

Il y a des textes sur le web qui sont ahurissants. En voici un qui provient du magazine Zataz. L'un des rédacteurs est suffisamment crédible pour être interviewé par le Nouvel Observateur. Le titre de cet article est «Google, Altavista et les autres …

Voici un extrait :

> *Etranges résultats*
> *« ... nous avons étudié les résultats de Google, et nous sommes tombés sur quelques interrogations : Premièrement ces recherches étant effectuées par un mot clé il ressort qu'il faut absolument l'avoir dans l'adresse de son site : ex "allo". Google nous renvoie sur des liens tels que allocine.fr, allocine.com, alloprof.qc.ca, allo.ch, … Il en va pour tout (sic) les mots sauf certains comme "cinéma" : où l'on obtient : allocine.fr, allocine.com, cinema.lu , fcm.fr, cine6.fr, … Vous aurez remarqué le lien fcm qui n'a rien à voir avec le cinéma...»*

La première remarque concernant « allo» est fausse. Google ne fait pas de «lemmatisation», il recherche le mot entier. Lorsque le mot n'est pas dans la page visible, il est soit dans la page source en métadonnée ou en spam - ex: même couleur pour la fonte et le fond, soit il est dans les pages référentes.

La seconde remarque est le cas de «cinéma». Pourquoi «le lien fcm n'a rien à voir avec le cinéma» ? Voici la réponse :

A la requête «cinéma», une unique adresse se glisse parmi les 10 premières références : http://www.fcm.fr. En cliquant sur cette adresse, nous nous rendons à http://www.fkgb.com/intro.html qui correspond à une agence de communication. Pourquoi ?

En cliquant sur le cache, Google restitue une page blanche puis apparaît la page du site de la société nommée FKGB. En fait, FKGB utilise une redirection. En sauvegardant la page originale, nous constatons que la page de fcm contient ce script :

```
<html>
<head>

 <script language="javascript">
<!--
document.location="http://www.fkgb.com";
-->
 </script>
 <title></title>
</head>
 <body>
<br>
 <br>
</body>
</html>
```

Nous observons que la balise titre n'enserre rien. Voilà pourquoi il n'y a aucun titre représentant cette page. De plus, la commande de redirection est mise dans un commentaire représenté par **<!– – – –>** qui n'est pas lu par les automates de recherche. La page est blanche pour eux.

Maintenant, pourquoi cette adresse est-elle parmi les premières ? Apparemment, il n'y a aucune raison. La réponse nous est fournie par AlltheWeb.

A la même requête «cinéma», l'adresse http://www.fcm.fr se trouve au-delà de la quatre-vingtième place. Voici la fiche que nous donne AlltheWeb:

> *www.fcm.fr*
> *http://www.fcm.fr/*
> *Find all 1,355 external web pages that link to "http://www.fcm.fr/"*
> *Find all 123 external web pages that contain the*

term "http://www.fcm.fr/"
Find all 1 web pages indexed under "http://www.fcm.fr/"
Discover who owns "fcm.fr"
See how "www.fcm.fr" used to look

La dernière indexation de AlltheWeb date du 4 décembre 2001. Si nous regardons les archives de Wayback Machine, nous nous rendons compte que la dernière date du site archivée est le 21 février 1999. Le site est au nom de France Cinéma Multimédia. Voilà où se trouve le mot «cinéma». La date d'archivage suivante est le 11 mai 2000. La page archivée est unique au site. Elle ne fait plus référence à France Cinéma Multimédia. La page est composée d'affiches publicitaires de films avec des hyperliens dans lesquels apparaissent le mot « cinéma». Le 18 mai, la page est modifiée pour des nouvelles affiches de film. Les lettres FKGB apparaissent dans le coin inférieur droit en minuscules avec des bandeaux publicitaires de films mais le mot «cinéma» est absent. Le 17 mai 2001, FKGB prend le contrôle de l'adresse. La firme de communication change la page d'accueil pour celle de redirection. Les automates sont leurrés.

Si nous regardons la fiche du site, nous avons le nom de Florent Darrault. En faisant une recherche sur le nom, nous apprenons que Florent Darrault a été président de FCM, France Cinéma Multimédia. FCM a été créé par M. Thierry Lhermitte, une agence de communication spécialisée dans le cinéma, soit FKGB, et M. Darrault, ancien employé de FKGB.
(Source : http://www.france.com/mag/interviews/) . La fermeture du site FCM correspond à l'éclatement de la bulle Internet. Le lien entre FCM et FKGB est évident.

Sans le cache, nous ne saurons jamais ce qui s'est passé. Il est cependant certain que le site répond aux mots «France», « cinéma» et « multimédia». A partir de là, les supputations commencent. Nous savons que les automates recherchent les mots dans les pages référentes et ceux hyperliés à des pages web externes sont prépondérants. Ensuite entre en jeu les algorithmes.

L'algorithme de Google met plus d'importance aux mots hyperliés et au nombre de pages référentes quel que soit le contenu. C'est ce que nous apprend la présence de la page blanche de fcm. La corrélation entre le mot et l'adresse de la page modifie la pondération. Ainsi, le lien «France Cinéma Multimédia» sera plus important que le lien «FCM». C'est pourquoi il est propulsé parmi les dix premières références en considérant que le nom complet est plus utilisé que le sigle.

AlltheWeb fait un équilibre entre les mots hyperliés des pages référentes et le contenu de la page. C'est pourquoi la page de FCM est loin. Le fait que cette page n'ait aucun texte visible montre bien que les algorithmes de recherche aient encore besoin d'être affinés pour retirer les bois morts.

Revenons à Google. L'adresse http://www.fcm.fr est citée dans 276 pages. La barre d'outils de Google nous informe que cette page a une PageRank de 6/10, celui de FKGB de 5/10.

Voici un tableau des dix premières références dans la recherche de «cinéma» avec Google France.

Allociné	8/10	5 330 liens
Cinéma.ca	6/10	462 liens
Cinéma Montréal	5/10	1660 liens
Cinema.lu	6/10	1600 liens
fcm (France Cinéma Multimédia)	6/10	276 liens
L'art du cinéma	6/10	370 liens
Nova	6/10	116 liens
Objectif Cinema	5/10	522 liens
Homecinema	5/10	266 liens
France Cinéma	5/10	88 liens

Il est à noter que le classement de Google est inconstant même à l'intérieur de 15 minutes. Les pages web peuvent changer de rangs. Certaines disparaîtront des dix premières références pour y revenir.

Conclusion

Si la page de fcm est parmi les premières références, c'est en raison de la qualité de ses pages référentes. FKGB utilise ce levier pour se hisser dans les premières références pour le mot «cinéma». Plusieurs sociétés utilisent cette tactique pour profiter de l'effet de levier. La crise des « point-com» de 2000 a fait sombrer plusieurs sites d'importances tel que le défunt répertoire canadien Carrefour avec le domaine « carrefour.net». Aujourd'hui, le site n'est plus référencé ni dans Google, ni dans AlltheWeb. Pourtant, des traces subsistent. Sur Google, Carrefour.net est lié par 612 pages et sur AlltheWeb par 14, 649 pages. Il est probable que la page ayant l'adresse www.carrefour.net puisse arriver en première page avec le mot «carrefour» sur Google.

Cet exemple démontre la vulnérabilité de PageRank de Google, la politique de mise à jour de Google pour des pages inchangées, la négligence des mises à jour des pages référentes et enfin les démonstrations fallacieuses qu'il induit. La société Google est consciente de cette vulnérabilité en voulant résoudre ce problème. La solution trouvée est un croisement entre les noms de domaines expirés et leur banque de données. Les changements sont prévus au courant de l'année. La suite sera pour le prochain numéro!

M.T. (Page consultée le 10 mars 2003). «Google, AltaVista et les autres.» [En linge]. In Zataz Magazine, 22 janvier 2001, revue 1er février 2003. http://www.zataz.com/zatazv7/mateur.htm

Clément Boulle (Page consultée le 10 mars 2003). «3 questions à... Damien Bancal. Google est un énorme «sniffeur»». [En ligne]. Le nouvel observateur, 6 mars 2003. http://permanent.nouvelobs.com/multimedia/20030306.OBS7818.html

Anonyme. (Page consultée le 10 mars 2003). «Interview: Acteur, scénariste et webmestre». [En ligne]. France Mag, [10 février 1997] http://www.france.com/mag/interviews/index.html

GoogleGuy. (Page consultée le 10 mars 2003). «Google Gives Good News About Expired Domains»[En ligne]. Webmaster World, 9 mars 2003. Rubrique «Google News» http://www.webmasterworld.com/

Marc Duval
Bibliothécaire professionnel
Service de recherche documentaire DSI
1280-4, bd de Montarville.
Longueuil. Québec. Canada
J4B 8B4
marcduval@dsi-info.ca
http://www.dsi-info.ca