V^E COLLOQUE TIC & TERRITOIRE : QUELS DEVELOPPEMENTS ?
Université de Franche Comté, Besançon, 9-10 juin 2006

ETAT DES LIEUX DES RESULTATS D'UNE RECHERCHE D'INFORMATION SIMULTANEE SUR LE MOTEUR DE RECHERCHE GOOGLE

Philippe PINCZON du SEL, pinczon@univ-tln.fr
Laboratoire I3M-EA 3820, Université du Sud Toulon-Var,
83957 La Garde Cedex, France

Mots clefs : Veille stratégique sur Internet, aide à la prise de décision, algorithme user centric, géolocalisation, classement de requêtes, pertinence de l'information, moteurs de recherche, navigation anonyme

Keywords: Competitive watch, decision-making, user centric algorithm, geolocalisation, queries rating, information relevance, search engines, anonymous web surfing

Résumé: Les moteurs de recherche s'appuient sur les habitudes de navigation des internautes et leur lieu de connexion afin d'affiner la pertinence des résultats. Les personnes en charge des campagnes de veille sur Internet pourraient influencer les résultats des requêtes : elles auraient été différentes selon la personne et le lieu.

Abstract: Search engines rely on the websurfers' private data to improve the revelance of their results. Thus, those who perform competitive watches on the Internet may influence these results: they would have been different if someone else would have done it from elsewhere.

1 Introduction

Lors de recherches d'informations sur les moteurs de recherche, notamment Google, les résultats d'une requête varient en nombre de réponses et en en pertinence non seulement selon le moment où elle est posée mais également selon la machine utilisée, où plus précisément l'endroit où elle se situe.

Les moteurs de recherche classent leurs résultats selon des critères de pertinence définis par des algorithmes. Ces derniers se basent sur le contenu des pages (content centric), sur leur popularité (link centric), sur leurs relations avec d'autres pages (algorithmes relationnels et contextuels) mais aussi sur des informations stockées sur les ordinateurs des internautes (user centric).

Or, comme tout traitement automatisé, l'utilisation d'algorithmes mathématiques peut entraîner des résultats différents si les informations sur lesquelles ils se fondent sont différentes. Ainsi, dans un précédent article (Pinczon du Sel, 2006), nous avons voulu démontrer qu'un veilleur peut influencer indirectement les résultats d'une veille Internet dont il a la charge : bien que deux ordinateurs strictement identiques aient été utilisés (l'un étant le clone de l'autre), nous avons obtenu des résultats différents aux requêtes soumises au moteur de recherche Google, car nous avions simulé deux types de connexion Internet différentes. L'algorithme « user centric » du moteur de recherche, trompé sur les données « utilisateur » d'un des deux ordinateurs, a fourni des informations différentes.

Cet article propose de poursuivre cette première expérience en proposant d'effectuer une veille Internet simultanée depuis plusieurs points géographiques situés en France. Ainsi, nous allons vérifier l'influence que peuvent avoir des personnes chargées d'une veille sur Internet sur ces résultats. Nous savons que les résultats d'une requête posée un jour sont différents le lendemain. Nous devrions désormais tenir compte également de leur position géographique ?

2 Le contexte de la recherche

2.1 Risques et Internet

Les réseaux apportent un enrichissement formidable des possibilités de faire et de mal faire, grâce notamment à l'interactivité, et surtout à la possibilité d'une circulation internationale des informations (Conseil d'Etat, 1998). Deux nouveaux types d'atteintes apparaissent selon les différents types de services proposés à l'utilisateur : soit les renseignements sont collectés sur les individus à travers des traitements visibles, soit une collecte d'informations est réalisée à l'insu de ceux-ci, au moyen de techniques mises en œuvre de façon cachée. Ainsi, dans quelques années, les technologies de tracing permettront aux serveurs Internet de garder votre identifiant et ainsi retrouver votre identité (Romagni, 1998).

Par ailleurs, la sécurité sur Internet peut être catégorisée en quatre grands thèmes : les attaques, les virus, le cryptage et l'espionnage (Revelli, 2000). Si les deux premières peuvent être associées au piratage en général, les dernières correspondent bien à la nouvelle forme de risque lié à Internet. Ainsi, l'auteur préconise-t-il de faire attention aux interventions dans les forums de discussion, d'éviter d'utiliser le réseau

d'une entreprise pour des recherches confidentielles et de désactiver les cookies à tout moment.

Le lien avec l'intelligence économique est donc établi : la méfiance des spécialistes de l'intelligence économique vis-à-vis de l'Internet tient à un certain nombre de risques du système : confidentialité difficile à respecter, désinformation possible, modification de documents, perte totale ou partielle de documents en cours de transfert, virus… Le spécialiste de la veille stratégique ou de l'intelligence économique prend conscience, en surveillant les autres, les concurrents, de la nécessité d'être discret, d'être prudent, de se protéger contre les risques et menaces diverses qui pourraient porter atteinte à son patrimoine informatif (Jakobiak, 2001).

Une première distinction est ici faite entre le risque de piratage proprement dit et la nécessité de se protéger contre les nouvelles technologies de l'Internet (NTI), puis une seconde distinction est proposée au sein même des NTI entre les informations diffusées en connaissance de cause par les internautes et celles qui leur sont soutirées sans qu'ils en soient avertis.

Divers éléments sur les internautes peuvent être recueillis de façon transparente sur le réseau : pratiques observées dans les forums de discussion, exploitation des messages électroniques et des annuaires. Mais au-delà des données nominatives circulant sur le réseau, il en est d'autres que l'utilisateur ne peut directement appréhender, alors que leur valeur informationnelle et les risques qu'elles représentent pour la vie privée des personnes sont importants : fichiers logs et cookies (Conseil d'Etat, 1998).

Les premiers servent à établir la communication entre ordinateurs distants. Les informations qui y sont stockées concernent d'une part les adresses des machines du réseau, dites adresses IP et en particulier celles de l'émetteur d'un message et de son destinataire, adresses auxquelles sont associées la date et l'heure de la connexion, des informations techniques caractérisant le type d'usage (accès au web, messagerie…) et d'autre part la requête (page du site que l'utilisateur veut visiter) ou le message proprement dit. Ces données sont collectées automatiquement par les fournisseurs d'accès et consignées dans un fichier dénommé « fichier log » (Conseil d'Etat, 1998).

Le cookie est un petit fichier texte émis par un serveur consulté par un utilisateur et enregistré sur le disque dur de celui-ci. Il comporte, en général, une date de validité et peut contenir l'information qu'aura souhaité y inclure le site visité (Conseil d'Etat, 1998) : ce peuvent être les dates et heures des visites (ou un compteur de visites), un numéro client, des réponses à un questionnaire, des mots-clés tapés dans le champ de recherche du site, des préférences ou le choix d'affichage et de navigation... Dans le cas d'un site web d'achats en ligne, les produits placés dans le « caddie » le sont en fait dans le cookie du site. En somme, le principe du cookie est purement marketing et a pour but de savoir si c'est la première fois que vous vous connectez à un certain site ou si au contraire vous y êtes déjà passé (Revelli, 2000) : il permet au responsable du site de mémoriser les précédentes consultations du site par l'internaute afin, soit de faciliter l'ergonomie de la visite, soit d'adapter les pages du site au « profil » de l'internaute tel qu'il est précisément déduit des « traces » conservées lors des précédentes visites (Conseil d'Etat, 1998).

Les internautes n'ont aucun contrôle sur ces fichiers logs, et les supprimer entraînerait automatiquement la perte de communication avec l'Internet. Par ailleurs, refuser d'accepter des cookies entrave considérablement la navigation sur les sites web : certains d'entre eux refusent

en effet leur accès dès lors que les cookies ont été désactivés.

2.2 L'hypothèse du risque pour l'intelligence économique

Nous le comprenons bien, afin de naviguer sur Internet, pour y mener une campagne de veille par exemple, toute personne est fichée, « profilée », et les informations qu'elle obtiendra pourront découler de ces informations.

A partir de ce constat, nous pensons que, pour toute campagne de veille sur Internet, nous obtenons des résultats différents selon la personnalité du ou des veilleurs ainsi que de leur situation géographique. Les décisions prises à la suite de ces campagnes peuvent donc à posteriori être également différentes.

Afin de vérifier cette hypothèse, nous avons mené une première expérience (Pinczon du Sel, 2006). Son but était de démontrer qu'à partir de deux postes strictement identiques, nous pouvions obtenir des résultats différents aux requêtes posées simultanément à des sites web si nous faussions les informations d'un des deux ordinateurs.

Ainsi, le moteur de recherche Google et le site d'enchères en ligne eBay renvoyaient effectivement des informations différentes aux deux postes informatiques. Les fichiers logs, les cookies et l'adresse IP ont donc bien joué un rôle prépondérant dans l'affichage des résultats.

Or, lors de cette expérience, les fichiers d'informations personnelles étaient strictement identiques, puisque les ordinateurs étaient des clones. Ayant été simplement inhibés sur l'un des ordinateurs, il nous fût impossible de connaître l'ampleur réelle de la différence de résultats entre deux postes distants ayant chacun leurs propres fichiers. Nous ne pouvions que constater une différence de résultats non significative d'un contexte réel de veille sur Internet.

Nous avons donc décidé de poursuivre l'expérience, cette fois avec des postes informatiques disséminés partout en France, afin d'établir un état des lieux le plus proche possible de la réalité d'une recherche d'information simultanée sur le moteur de recherche Google.

3 Recherche d'information simultanée sur le moteur de recherche Google

3.1 Le protocole

Afin de réduire autant que possible le risque d'erreur pouvant entraîner de faux résultats, nous avons établi le protocole suivant :

- Pour d'éviter les erreurs dues à l'espacement dans le temps des réponses, les différentes connexion à Internet doivent avoir lieu le même jour à la même heure. Le jour choisi fût le vendredi 10 juin 2005 à 14h00 ;

- Pour limiter les erreurs dues à l'utilisation d'un matériel spécifique, nous devons noter, pour chacun participant, le type de connexion (56k, adsl, intranet), le système d'exploitation (Mac, Linux, Windows) ainsi le navigateur Internet utilisés ;

- Par ailleurs, nous prévoyons que le seul langage utilisé par les navigateurs soit le français.

L'expérience proprement dite s'est ensuite déroulée en deux étapes :

- Une première connexion à Google France pour y taper une série de 10 mots-clés et enregistrer les 500

premiers résultats de chaque requête ;

- Une seconde connexion à Google.com, version « google in english », pour y effectuer une procédure identique, mais avec des mots-clés différents.

Les mots-clés ont été choisis directement sur le site de Google, d'après leur classement « Zeitgeist » qui liste chaque mois les mots-clés les plus populaires par langue. Nous avons pris les 10 premiers mots-clés des langues française du mois d'avril 2005 (ceux du mois précédent l'expérience n'étant pas encore accessibles) et anglaise du mois précédent l'expérience (voir tableau 1).

3.2 L'expérience

35 élèves furent conviés à participer à l'expérience, dont certains à l'étranger, notamment au Canada et en Australie.

	http://www.google.fr	http://wwwgoogle.com (option « google in english »)
Position	**Google Zeitgeist (avril 2005)**	**Google Zeitgeist (mai 2005)**
1	meetic	nintendo revolution
2	impots	kasey kahne
3	brice de nice	ps3
4	manga	star wars episode 3
5	jeux	e3
6	tiscali	xbox 360
7	sfr	kylie minogue
8	caf	natalie portman
9	pape	preakness
10	pmu	the bachelor

Tableau 1 : classement « Zeitgeist » de Google pour les langues française et anglaise des mois d'avril et de mai 2005

10 ont effectivement répondu à l'appel, tous en France. Le tableau 2 nous renseigne sur leurs lieux de connexion à Internet ainsi que les dates et horaires auxquels ils ont effectué les recherches et leur configuration matérielle et logicielle.

	Lieu	**OS / navigateur**	**Date (Heure)**
Elève 1	83500 La Seyne/Mer	XP / IE6	10/06/05 (14:20)
Elève 2	39000 Lons-le-Saunier	XP / FireFox 1.0.4	10/06/05 (14:00)
Elève 3	83130 La Garde	XP / FireFox 1.0.4	10/06/05 (14:00)
Elève 4	83210 Solliès-Toucas	XP / FireFox 1.0.4	10/06/05 (14:00)
Elève 5	75001 Paris	XP / FireFox 1.0.3	10/06/05 (14:00)
Elève 6	13000 Marseille	XP / IE6	10/06/05 (14:15)
Elève 7	83130 La Garde	XP / Opera 8.0	10/06/05 (14:00)
Elève 8	83200 Toulon-Ouest	(N.C.)	11/06/05
Elève 9	13400 Aubagne	XP / FireFox 1.0.4	12/06/05
Elève 10	83130 La Garde	XP / IE6	(N.C.)

Tableau 2 : lieux et dates de connexion des élèves ayant participé à l'expérience

Les résultats de quatre étudiants furent écartés pour différentes raisons :

- Connexion non effectuée le jour demandé (élèves 8 et 9) ;
- Connexion strictement identique (même lieu de stage) qu'un autre participant (élèves 3, 7 et 10 – nous n'avons conservé que les résultats de l'élève 3) ;
- Insuffisance de renseignements (élèves 8 et 10).

En outre, les six participants retenus ont l'avantage d'avoir utilisé pour leurs recherches du matériel et des configurations logicielles très proches.

3.2.1 Un nombre de résultats différents

Les tableaux 3 et 4 présentent les résultats de chacun des participants retenus. Afin de mieux comprendre les différences de résultats, nous avons noté les adresses IP des Google Data Centers (serveurs) qui ont répondu à chaque requête.

A la lecture de ces tableaux, nous constatons plusieurs choses :

Nous constatons une première différence entre le lieu géographique de la connexion et le Google Data Center ayant répondu. Les Google Data Center ne répondent donc pas par région, mais de façon aléatoire, certainement en fonction de la disponibilité du réseau.

Ainsi, le Google Data Center ayant répondu aux requêtes de l'élève en région Parisienne a également répondu à celui de Marseille. En revanche, deux Google Data Center différents ont répondu aux élèves en stage à La Garde et à Toulon, villes distantes d'une vingtaine de kilomètres.

	Eleve 1 83500 La Seyne/Mer	**Eleve 2 39000 Lons-le-Saunier**	**Eleve 3 83130 La Garde**
Mots-clés	**Nombre de réponses**		
GOOGLE.COM	**64.233.183.104 (XP / IE6)**	**216.239.59.104 (XP / FireFox 1.0.4)**	**216.239.59.104 (XP / FireFox 1.0.4)**
the bachelor	33 800 000	46 800 000	45 800 000
e3	16 100 000	31 400 000	15 000 000
kasey kahne	785 000	1 170 000	1 340 000
kylie minogue	1 470 000	3 670 000	3 450 000
natalie portman	2 110 000	3 320 000	2 160 000
nintendo revolution	16 500	3 390 000	7 800 000
preakness	689 000	587 000	639 000
ps3	6 690 000	8 410 000	7 870 000
star wars episode 3	14 100 000	17 100 000	16 100 000
xbox 360	9 870 000	10 100 000	18 700 000
GOOGLE.FR	**64.233.183.104 (XP / FireFox 1.0.4)**	**66.102.9.104 (2000 / FireFox 1.0.4)**	**216.239.59.104 (XP / FireFox 1.0.4)**
brice de nice	929 000	628 000	607 000
caf	2 050 000	2 490 000	4 120 000
impots	1 780 000	3 440 000	3 500 000
jeux	13 400 000	54 000 000	10 600 000
manga	6 410 000	24 200 000	23 300 000
meetic	887 000	883 000	891 000
pape	1 730 000	3 650 000	3 290 000

pmu	864 000	609 000	606 000
sfr	1 760 000	4 820 000	4 820 000
tiscali	4 250 000	4 670 000	4 670 000

Tableau 3 : Nombre de réponses obtenues dans Google pour les élèves 1 à 3

Quelque soit la langue demandée, le même Google Data Center répond, sauf pour les élèves 2 et 6, très probablement en conséquence d'une indisponibilité temporaire des Data Centers. Seul l'élève 2 avait changé de configuration logicielle entre-temps.

Nous obtenons systématiquement un nombre de résultats différents selon les Google Data Center. En effet, presque aucun mot-clé n'affiche le même nombre de résultats dans deux Google Data Center différents, sauf pour les mots-clés « sfr » et « tiscali » qui obtiennent les mêmes résultats dans les Google Data Center 66.102.9.104 et 216.239.59.104.

L'amplitude de l'écart du nombre de résultats oscille entre de 0,89% pour le mot clé « meetic » et 99,79% pour « nintendo revolution », la moyenne pour les vingt mots-clés étant de 45,39% (voir tableau 5).

	Eleve 4 83210 Solliès-Toucas	**Eleve 5 75001 Paris**	**Eleve 6 13000 Marseille**
Mots-clés	**Nombre de réponses**		
GOOGLE.COM	**66.102.9.104 (XP / FireFox 1.0.4)**	**66.102.9.104 (XP / FireFox 1.0.3)**	**66.102.9.104 (XP / IE6)**
the bachelor	48 900 000	48 900 000	48 900 000
e3	34 300 000	34 300 000	34 300 000
kasey kahne	605 000	605 000	605 000
kylie minogue	4 020 000	4 020 000	4 020 000
natalie portman	3 070 000	3 070 000	3 070 000
nintendo revolution	8 140 000	8 140 000	8 140 000
preakness	573 000	573 000	573 000
ps3	7 990 000	7 990 000	7 990 000
star wars episode 3	22 400 000	22 400 000	22 400 000
xbox 360	19 000 000	19 000 000	19 000 000
GOOGLE.FR	**66.102.9.104 (XP / FireFox 1.0.4)**	**66.102.9.104 (XP / FireFox 1.0.3)**	**216.239.59.104 (XP / IE6)**
brice de nice	628 000	628 000	607 000
caf	2 490 000	2 490 000	4 120 000
impots	3 440 000	3 440 000	3 500 000
jeux	54 000 000	54 000 000	10 600 000
manga	24 200 000	24 200 000	23 300 000
meetic	883 000	883 000	891 000
pape	3 650 000	3 650 000	3 290 000
pmu	609 000	609 000	606 000
sfr	4 820 000	4 820 000	4 820 000
tiscali	4 670 000	4 670 000	4 670 000

Tableau 4 : Nombre de réponses obtenues dans Google pour les élèves 4 à 6

Au sein de chaque Google Data Center les résultats sont stables, excepté le Google Data Center 216.239.59.104 qui affiche des résultats différents pour les mots clés en langue anglaise. L'écart du nombre de résultats affichés varie alors de 2,13% à 56,53%, pour une moyenne entre les dix mots-clés de 23,08% (voir tableau 5).

Notons également que le Google Data Center 66.102.9.104, bien que très stable dans le nombre de ses réponses à chaque étudiant, a affiché des résultats globalement moindres deux jours plus tard. En effet, ce Google Data Center a alors répondu aux requêtes de l'élève d'Aubagne avec un taux de réponses variant de +34,09% à -80,18% par rapport aux résultats qu'il propose ici (voir tableau 6).

D'une façon générale, le Google Data Center 66.102.9.104 a fourni le plus de réponses, tandis que le Google Data Center 64.233.183.104 en a fourni le moins.

Mots-clés	Diff. en % entre tous Data Centers	Diff. en % (66.102.9.104)	Diff. en % (216.239.59.104)
the bachelor	30,88	0	2,13
e3	56,26	0	52,22
kasey kahne	54,85	0	12,68
kylie minogue	63,43	0	5,99
natalie portman	36,44	0	34,93
nintendo revolution	99,79	0	56,53
preakness	16,83	0	8,13
ps3	20,45	0	6,42
star wars episode 3	37,05	0	5,84
xbox 360	48,05	0	45,98
brice de nice	34,66	0	0
caf	50,24	0	0
impots	49,14	0	0
jeux	80,37	0	0
manga	73,51	0	0
meetic	0,89	0	0
pape	52,60	0	0
pmu	29,86	0	0
sfr	63,48	0	0
tiscali	8,99	0	0

Tableau 5 : Différence des résultats en pourcentage entre tous les Google Data Centers, puis au sein du même des Google Data Centers

Mots-clés	10/06/05 (Paris)	12/06/05 (Aubagne)	Diff. en %
the bachelor	48,900,000	32,900,000	-32,72
e3	34,300,000	15,700,000	-54,22
kasey kahne	605,000	732,000	+17,35
kylie minogue	4,020,000	1,390,000	-65,42
natalie portman	3,070,000	2,000,000	-34,85
nintendo revolution	8,140,000	3,080,000	-62,16
preakness	573,000	772,000	+25,77
ps3	7,990,000	6,550,000	-18,02
star wars episode 3	22,400,000	13,700,000	-38,84
xbox 360	19,000,000	9,130,000	-51,94
brice de nice	628 000	824 000	+23,78

caf	2 490 000	2 140 000	-14,05
impots	3 440 000	1 800 000	-47,67
jeux	54 000 000	10 700 000	-80,18
manga	24 200 000	6 530 000	-73,01
meetic	883 000	890 000	+0,78
pape	3 650 000	1 890 000	-48,22
pmu	609 000	924 000	+34,09
sfr	4 820 000	1 810 000	-62,45
tiscali	4 670 000	4 400 000	-5,78

Tableau 6 : Différence des résultats du Google Data Center 66.102.9.104 entre le 10 juin 2005 et le 12 juin 2005

3.2.2 Un classement des résultats différent

En ce qui concerne le classement des réponses on s'attend, à la vue du nombre de réponses différentes entre Google Data Centers ainsi qu'au sein même de chaque Google Data Center, à observer également des disparités, le Google Data Center 66.102.9.104 mis à part car resté stable.

En effet, le classement des réponses au sein du Google Data Center 66.102.9.104 reste le même pour tous les mots clés, en revanche, pour le Google Data Center 216.239.59.104 nous obtenons des changements de positions en moyenne à partir de la 24ème place, mais uniquement pour la langue anglaise : les mots-clés français demeurent inchangés (voir tableau 7).

Nous pouvons par ailleurs comparer ce tableau avec le tableau 5 : le Google Data Center 216.239.59.104 y affiche un nombre de réponses différent pour la langue anglaise seulement, ce qui semble confirmer les premiers résultats.

Mots-clés	66.102.9.104	216.239.59.104
the bachelor	Idem	32
e3	Idem	10
kasey kahne	Idem	30
kylie minogue	Idem	28
natalie portman	Idem	2
nintendo revolution	Idem	7
preakness	Idem	13
ps3	Idem	69
star wars episode 3	Idem	27
xbox 360	Idem	20
brice de nice	Idem	Idem
caf	Idem	Idem
impots	Idem	Idem
jeux	Idem	Idem
manga	Idem	Idem
meetic	Idem	Idem
pape	Idem	Idem
pmu	idem	Idem
sfr	Idem	Idem
tiscali	Idem	Idem

Tableau 7 : Position du premier changement de réponse au sein d'un même Google Data Center (100 premières réponses)

Si l'on considère le Google Data Center 66.102.9.104 comme référence, compte-tenu de sa stabilité et de son nombre de réponses globalement supérieur aux autres, nous observons énormément de disparités dans le classement des réponses entre Google Data Centers (voir tableau 8).

Un mot-clé seulement sur vingt, « pape », conserve le même classement quelque soit le Google Data Center, et six mots-clés ont un classement identique entre le Google Data Center 66.102.9.104 qui sert ici de référence et le Google Data Center 216.239.59.104.

Ces chiffres semblent confirmer les tableaux 3 et 4 qui précisent que le Google Data Center 64.233.183.104 est le moins performant des trois de cette analyse.

A l'inverse, dix-mots clés – soit la moitié d'entre-eux – voient leur classement modifié avant la 15ème place, tous Google Data Centers confondus.

Attention toutefois, cette analyse a été effectuée sur les cent premiers résultats de chaque requête. Le terme « Idem » entend ici qu'aucun changement n'a été constaté jusqu'à la centième place.

Mots-clés	216.239.59.104	64.233.183.104
the bachelor	28	25
e3	10	13
kasey kahne	22	19
kylie minogue	66	66
natalie portman	Idem	26
nintendo revolution	77	1
preakness	31	(N.C.)
ps3	22	2
star wars episode 3	9	10
xbox 360	Idem	13
brice de nice	8	6
caf	29	29
impots	Idem	29
jeux	65	63
manga	Idem	11
meetic	58	9
pape	Idem	Idem
pmu	Idem	20
sfr	5	5
tiscali	38	6

Tableau 8 : Position du premier changement de réponse par rapport au Google Data Center 66.102.9.104 (100 premières réponses)

3.3 Application de l'expérimentation à la veille

Les moteurs de recherche évoluent et tendent vers une amélioration de la pertinence des résultats proposés aux internautes. Or, cette pertinence est toute relative : elle est basée sur l'hypothèse que les résultats sont d'autant plus pertinents pour un internaute s'ils prennent en compte ses goûts, ses habitudes de navigation et sa situation géographique.

Nous pensons que la recherche d'informations sur Internet dans le cadre d'une veille technologique doit s'affranchir de ces évolutions : pour atteindre un niveau optimal d'efficacité, une veille sectorielle sur Internet doit demeurer impersonnelle, les résultats des recherches sont faussés si on y intègre la localisation géographique et les goûts personnels des veilleurs en charge de l'action.

Or, nous voyons bien ici les conséquences : quelle que soit la personne et le lieu, les résultats d'une requête au moteur de recherche Google sont différents en nombre et en classement. D'un Google Data Center à un autre et de façon simultanée, nous obtenons pour certains mots-clés 99,79% de réponses en plus et des classements qui changent dès la première place.

Nous pensons que ces faits sont à prendre en considération lors des campagnes de veille sur Internet, les veilleurs doivent être sensibilisés au fait qu'ils peuvent obtenir d'autres réponses s'ils se connectent directement aux différents Google Data Centers au lieu de se fier à la connexion automatique vers l'un ou l'autre Google Data Center.

4 Conclusion

Nous n'avons fait ici qu'un état des lieux d'une recherche d'information simultanée sur un moteur de recherche, nous n'avons pas encore appliqué cette expérience aux autres moteurs de recherche.

Par ailleurs, le but de cet article n'était pas de déterminer les raisons de ces disparités ; elles feront l'objet d'autres recherches par exemple en recommençant cette même expérience mais en demandant aux participants de se connecter à un Google Data Center en particulier. Nous pourrons ainsi mieux isoler le facteur de personnalisation de Google.

Enfin, nous n'avons pas fait ici d'analyse précise de la pertinence des résultats. Nous avons vu que les classements sont différents, à un moment t, selon les Google Data Centers, mais sont-ils si différents au point de menacer les prises de décisions qui en découleraient ?

5 Bibliographie

Conseil d'Etat. 1998. Internet et les réseaux numériques. La documentation française.

Jacobiak, François. 2001. L'intelligence économique en pratique. 2ème édition. Les éditions d'organisation.

Pinczon du Sel, Philippe, 2006, Influence des algorithmes des outils de recherche sur les requêtes d'une veille : exemple de la navigation anonyme. Colloque VSST Lille 2006.

Revelli, Carlos. 2000. Intelligence stratégique sur internet. Dunod.

Romagni, Patrick, et Wild, Véronique. 1998. L'intelligence économique au service de l'entreprise.