L'exploitation de l'age d'une page web : quelles perspectives pour l'analyse cybermétrique

Eric Boutin
boutin@univ-tln.fr

IUT de Toulon / Laboratoire Le Pont- USTV ★ BP 132 ★ F-83957 La Garde Cedex France+ 33 4 94 14 23 56

Mots clés :
Analyse cybermétrique, indicateurs temporels, veille sectorielle, indicateur de pertinence, moteur de recherche

Keywords :
Cybermetric analysis, temporal indicators, sectorial wakefullness, relevancy indicator, search engine

Palabras clave :
Análisis cibermétrico - indicadores temporales - vigilia sectorial - indicador de pertinencia - motor de búsqueda

Résumé :
L'information web ne comporte pas toujours de référent temporel. Rares sont les pages web qui portent une mention explicite de leur date de création, de leur date de mise à jour. Or, la validation d'une information quelle qu'elle soit (information scientifique, technique, économique, stratégique) passe par l'ancrage de cette information dans le temps et l'espace.
Partant de ce constat, l'objectif de cette communication est d'estimer la date de création d'une page web et de montrer l'intérêt que présente la notion d'âge d'une page web pour l'analyse cybermétrique. La question de la détermination de l'âge d'une page web sera abordée dans un premier temps. Dans un second temps, nous traiterons de l'impact de la disposition d'une information temporelle sur les indicateurs cybermétriques au niveau microscopique et macroscopique. Deux validations expérimentales serviront de guide à cette étude d'impact.

1. Introduction :

Le web est construit selon une logique de substitution non sédimentaire : de nouvelles pages portant le même nom que les anciennes remplacent leur contenu. L'internaute dispose ainsi d'un panorama hétéroclite de pages web où les informations récentes côtoient des informations plus anciennes sans que cette distinction soit toujours possible.

Une autre caractéristique de l'information web est en effet de ne pas forcément comporter de référent temporel. Rares sont les pages web qui portent une mention explicite de leur date de création, de leur date de mise à jour. Or, la validation d'une information quelle qu'elle soit (information scientifique, technique, économique, stratégique) passe bien souvent par l'ancrage de cette information dans un référent spatial et temporel. Ces règles d'usage sont malmenées dans le cas de l'information sur internet.

Partant de ce constat, l'objectif de cette communication est d'affecter un âge à une page web et d'utiliser cet indicateur à divers niveaux. La question de la détermination de l'âge d'une page web sera abordée dans un premier temps en même temps que la question des limites de cette affectation. Dans un second temps, nous traiterons de l'impact de la disposition d'une information temporelle sur les indicateurs cybermétriques. Pour des raisons didactiques, cet impact sera examiné à deux niveaux : au niveau microscopique et au niveau macroscopique.

2. La recherche d'informations temporelles sur une page web

Différentes informations temporelles peuvent caractériser une page web. Nous en avons identifié trois :

2.1. Date de dépôt du nom de domaine du site sur lequel est située la page

Il s'agit d'une information collectée par les registrers lors du dépôt d'un nom de domaine et mise à la disposition des internautes à travers l'utilisation d'une commande whois sous unix. L'exemple présenté figure 1 donne un exemple d'information renvoyée par un registrer pour le site encyclopedia.com.

```
<whois domain="encyclopedia.com">
 [whois.godaddy.com]
Registrant:
   Alacritude, LLC
   590 N. Gulph Rd
   King of Prussia, Pennsylvania 19406
   United States
   Domain Name: ENCYCLOPEDIA.COM
     Created on: 23-Jan-98
     Expires on: 22-Jan-05
     Last Updated on: 11-Dec-03
   Administrative Contact:
     Admin, Domain  domain-admin@alacritude.com
</whois>
```

Figure1 : exemple de données obtenues par une requête whois

En exploitant les informations contenues dans ce fichier, on peut récupérer la date de dépôt du nom de domaine et différentes informations identitaires sur son déposant.

L'utilisation de cette information soulève plusieurs problèmes :

- Un problème de précision : La date de dépôt du nom de domaine ne correspond pas forcément à la date à laquelle le site web a été activé et ce pour différentes raisons :
    - Le nom de domaine a pu être déposé pour réserver le nom avant même que les pages du site soient disponibles sur internet

    - Une version antérieure du site a pu être déposée avec une autre adresse.
- Un problème juridique puisqu'il est interdit de faire un usage systématique ou marchand des informations récupérées sur les registrers. La collecte automatique par une routine de l'information des registers se heurte de ce fait à certains obstacles : certains registers autorisent le requétage successif de leur base après un certain laps de temps ce qui est bloquant pour une analyse d'un grand nombre de données.
- Un problème de pertinence dans le cas de site communautaire par exemple : en effet si le site web est déposé dans un espace communautaire, la date de dépôt du nom de domaine sera la même pour tous les sous domaines hébergés par ce site communautaire.
- Un problème de formatage des données renvoyées par les registers : chaque register renvoie l'information correspondant au nom de domaine demandé en suivant une mise en forme spécifique qui ne présente aucun caractère homogène. De plus, ce formatage des données n'est pas toujours très explicite. A ce titre, les fichiers renvoyés par la Frnic sont peu exploitables par des routines automatiques pour extraire la date de dépôt du site web.
- Un problème de disponibilité : Parfois, il est impossible d'obtenir les informations du register soit parce qu'on l'a interrogé à des intervalles de temps trop contigus soit parce que le déposant du nom de domaine a fait en sorte qu'on ne puisse pas accéder à cette information.

Les expérimentations que nous avons réalisées nous ont conduit à estimer que la date de création du site web pouvait être exploitée avec pertinence dans environ 50% des sites web. Il est donc délicat de prétendre conduire une analyse exhaustive à partir de cette information.

2.2. Date de dernière mise à jour de la page sur le web.

Cette information est laissée à la discrétion du concepteur du site web : dans la pratique elle est peu renseignée. Etant donné que cette information est utilisée dans les algorithmes des moteurs de recherche pour qualifier la fraîcheur d'une page web, certains sites peuvent être rafraîchis automatiquement pour augmenter leur pertinence sur les moteurs si bien que cette information perd de son intérêt. De plus, avec le développement du web dynamique, les pages web sont créées à la demande de l'internaute. La date de création de la page se confond alors avec la date de la requête et n'a donc plus de sens. Ces trois raisons nous ont conduit à ne pas exploiter cette information.

2.3. Date de dépôt de la page sur le web.

Cette information peut est approchée par trois dates qui lui servent de majorants et de minorants :

- la date de dépôt d'une page web sur le web est au moins supérieure à la date de dépôt du nom de domaine.
- la date de dépôt d'une page sur le web est antérieure à la date correspondant à la date de mise à jour la plus ancienne des pages qui pointent vers la page à définir. Ce critère hérite des limites du critère précédant.
- Le recours au site web.archive.org pour connaître la première mise à disposition d'une page sur le web. Ce moteur de recherche est le fruit d'un travail d'archivage du web [8] afin, selon Feise [4] d'en garder la mémoire. Ainsi le moteur de recherche dispose de robots qui ont scanné le web à divers moments de son histoire et ont conservé des images fidèles à intervalles réguliers. Il est ainsi possible de récupérer ces différentes versions de pages web pour les analyser. L'utilisation de l'interface de web.archive.org est assez intuitive. Lorsqu'on tape l'adresse web d'une page web, on obtient un tableau qui renvoie une liste de dates rangées par année. Il s'agit des différentes instances d'une page stockée. La date la plus ancienne correspond à la date à laquelle le moteur a scanné la page pour la première fois. Cette date est postérieure à la date de mise à disposition de la page sur le web. Pour pouvoir utiliser cet outil, il faut s'assurer que l'écart entre la date de création de la page web et celle de son référencement dans web.archive.org est limitée et que ce moteur a une couverture satisfaisante du web.

3. La date d'une page web : un moyen d'affiner l'indicateur de pertinence d'un moteur de recherche

Les indicateurs de pertinence des moteurs de recherche utilisent plusieurs technologies mêlant analyse textuelle et étude du contexte relationnel de cette page. Toutefois, depuis Google [7] , les indicateurs de pertinence sont basés essentiellement sur l'analyse des liens entrants sur une page. Plus une page reçoit de liens émanant de pages pertinentes et plus cette page devient pertinente en héritant d'une partie de la notoriété des pages qui la citent.
Les indicateurs relationnels ont, du fait de ces choix technologiques, tendance à valoriser les pages les plus anciennes et donc à favoriser une certaine inertie comme l'ont montré Fetterly et al [5] et Lim et al [9]. En effet la probabilité pour une page d'être citée par d'autres pages sera d'autant plus forte que la page est plus ancienne. Les moteurs de recherche privilégiant l'analyse relationnelle favorisent donc les pages anciennes et par corollaire pénalisent les pages émergentes comme le soulignent Ntoulas et al [6]. Si on se place dans la problématique de la veille stratégique, un des objectifs est d'identifier une information la plus en amont possible de son processus de diffusion. Pour pouvoir utiliser le résultat des outils de recherche généraliste dans un processus de veille, nous proposons donc d'apporter un patch correctif qui a pour objet d'analyser l'information relationnelle d'une page en même temps que l'âge de cette page. L'objectif est de permettre à des pages récentes peu citées mais proportionnellement plus citées que des pages plus anciennes de figurer en bonne place dans les résultats du moteur. Nous présenterons dans un premier temps le principe général de la méthode que nous avons retenue. Nous présenterons dans un second temps, quelques résultats expérimentaux.

3.1. Principe général de la méthode

3.1.1. Le choix de l'hypothèse de départ

Nous avons choisi de mesurer l'âge d'une page web par le temps écoulé depuis la première fois que cette page a été référencée dans le moteur web.archive.org. Nous avons choisi de mesurer le nombre de liens entrants d'une page web par le nombre de résultats renvoyés par la commande link adressée à Google.
Toute la question reste alors de savoir de quelle façon l'âge de la page web va être pris en compte pour nuancer l'indicateur relationnel. Deux optiques, reposant chacune sur une hypothèse spécifique, ont été investiguées successivement :
- hypothèse 1 : la création de liens entrants sur une page web est continue au fil du temps. Dans ce cas, la correction pourrait être envisagée en divisant l'indicateur relationnel pur du moteur par l'ancienneté de la page web. Toutefois, un tel indicateur revient à considérer que le processus de création de liens entrants sur une page est un processus linéaire ce qui ne correspond pas à la réalité.
- hypothèse 2 : la création de liens entrants sur une page web n'obéit pas à un processus linéaire. Cette hypothèse revient à affiner le modèle précédant et est beaucoup plus proche de la réalité. En effet, on a pu observer que le processus de création de liens à partir d'une page web s'apparente davantage à un processus non linéaire. L'étude expérimentale que nous avons conduite nous a permis de modéliser le processus de création de liens entrants sur une page web depuis sa date de mise sur le web jusqu'à un instant t. Cette modélisation permet de connaître la fonction mathématique de création de liens entrants sur une page en fonction de son âge et de corriger les indicateurs relationnels du moteur.

3.1.2. Modélisation du nombre de liens entrants sur une page web en fonction de son âge :

L'hypothèse que nous faisons est que le nombre de liens entrants sur une page web est en relation non linéaire avec l'âge de cette page web. Cette relation peut être représentée de façon simplifiée par la courbe présentée Figure 2.

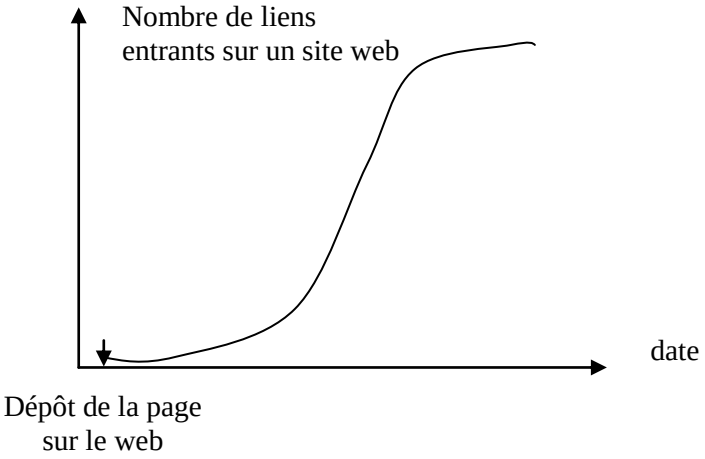

Figure 2 : courbe exprimant la relation entre l' âge d'une page web et son nombre de liens entrants

- Toutefois, l'âge d'une page web n'est pas le seul déterminant de la création de liens entrants sur une page web si bien qu'il y a de fortes variations autour de cette configuration idéale. Certains sites, ayant bien compris le fonctionnement des indicateurs de pertinence des outils de recherche, ont des stratégies de développement de liens entrants.
- D'autre part, le nombre de liens entrants sera proportionnellement plus fort dans certains domaines particuliers.

Notre méthode se propose de gommer l'effet âge mais sera sans effet sur les autres caractérisant de la popularité d'une page.
Conscients du caractère très contingent du processus de création de liens entrants pour une page, nous avons choisi, pour limiter un peu cette contingence, de travailler dans un contexte donné. A partir d'une requête adressée au moteur de recherche, on récupère un certain nombre de pages web. Celles-ci traitant d'un même sujet, on peut faire l'hypothèse que ceteris paribus la stratégie de création de liens entrants est la même pour ce domaine. Cette hypothèse constitue à elle seule une limite de la méthode. En effet, pour un même ensemble de résultats renvoyés par un moteur, on peut avoir plusieurs groupes de sites web ayant des stratégies de liens entrants fort différentes. Si par exemple, on tape « Provence alpes cote d'azur » sur Google, on aura d'une part des sites institutionnels qui sont dans une dynamique relationnelle particulière et des sites touristiques marchands qui sont dans une autre dynamique relationnelle.

3.2. Un exemple pour illustrer la démarche

Pour connaître la dynamique de création de liens entrants pour un ensemble de pages web, deux méthodes expérimentales sont envisageables.
La première consiste à suivre tout au long de leur vie un certain nombre de pages web en recueillant à divers instants des informations relatives aux liens entrants sur ces pages web. C'est la méthode retenue par Ntoulas et al [4]. Une telle démarche nécessite un recul important.
Une autre démarche consiste à analyser le nombre de liens entrants à un instant t d'un échantillon de pages web sur un sujet donné, chaque page web analysée ayant un âge. C'est cette solution que nous avons privilégiée.

On a considéré les 1000 premières réponses disponibles sous Google pour cette requête.

On s'est intéressé dans cette expérimentation à la requête « intelligence économique » qui correspond à un champ que nous connaissons particulièrement. Cette requête a été adressée au moteur Google. Nous avons considéré un ensemble primaire de pages nécessaire pour constituer un sous ensemble de

139 pages référencées dans le moteur web.archive.org. Il est à noter que web.archive.org référence globalement 60% environ des 300 premières pages renvoyées par Google pour cette requête. Pour chacune de ces 139 pages issues du résultat de Google, nous avons récupéré la date de mise à disposition de cette page sur internet (ou du moins dans web.archive.org) ainsi que le nombre de liens entrants sur chacune d'entre elles (à partir de la commande link : tapée sur google). La représentation de ces deux dimensions (âge de la page, nombre de liens entrants) fait l'objet du graphe représenté figure 3

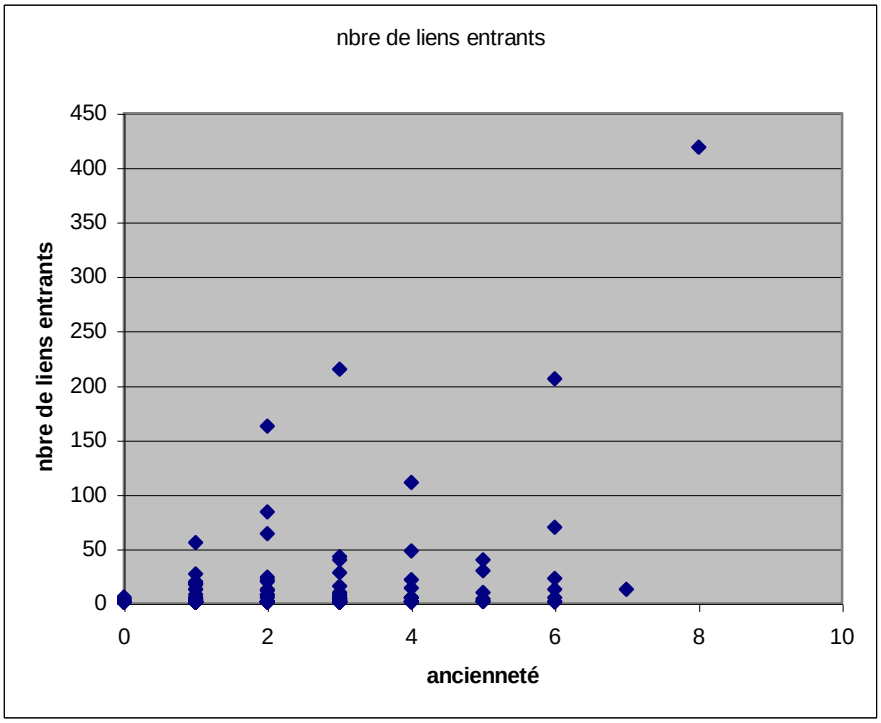

Figure 3 : Relation entre ancienneté d'une page web et popularité de cette page

Cette même information peut être représentée sur un graphique semi logarithmique dans lequel on représente en ordonnées non pas le nombre de liens entrants mais le logarithme décimal de cette valeur. Pour pouvoir réaliser cette opération, nous avons rajouté un lien entrant à chaque page, le logarithme n'étant défini que pour des valeurs non nulles.

Le graphique semi logarithmique est donné dans la figure 4

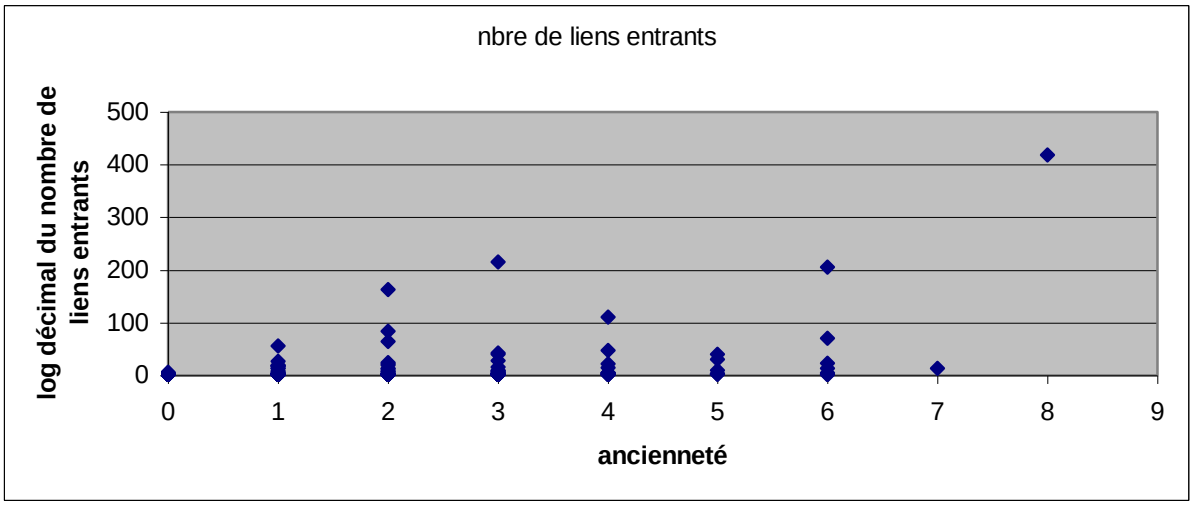

Figure 4 : Relation entre ancienneté d'une page web et popularité de cette page

Les données de ce graphique sont fortement dispersées. Toutefois, l'évolution générale du nuage de points met en évidence un trend légèrement ascendant qui tendrait à valider notre hypothèse de départ. Plus une page est ancienne, plus, ceteris paribus, elle a de chance de recevoir un nombre significatif de liens entrants. Pour obtenir plus précisément les caractéristiques de ce trend, nous avons procédé, sur la représentation semi logarithmique, à un ajustement linéaire.
Cette droite d'ajustement linéaire a les coefficients suivants :
a= 0.1408
b= 0.2601
Compte tenu de la forte dispersion des valeurs, le coefficient de corrélation associé est faible (0.387)
La relation suivante peut donc être proposée

LOG(nombre de liens entrants)=0.1408 âge + 0.2601 ou nombre de liens entrants $= 10^{(0.1408 Age + 0.2601)}$

Cette équation nous a permis d'estimer pour chacune des 139 pages web un nombre de liens entrants théoriques qu'elle aurait du recevoir compte tenu de son âge.
Prenons l'exemple de la page veille.com qui ressort en première position sur Google. Cette page a été référencée dans web.archive.org le 6/12/1998 soit il y a 6 ans. Cette page reçoit 205 liens entrants. Compte tenu de l'ancienneté de cette page, celle-ci, si on retient les paramètres définis précédemment aurait du recevoir 12,73 liens entrants. Elle reçoit donc proportionnellement plus de liens que ce auquel elle aurait pu prétendre.
Comparons maintenant la popularité de 2 pages :
La page www.veille.com/labo.htm reçoit 4 liens entrants tout comme la page www.i-km.com/colloque_europeen_ie.htm . Toutefois la première page a 6 ans d'existence alors que la seconde n'a pas un an.
Le rapport entre le nombre de liens entrants observé et le nombre de liens entrants théorique fournit un classement qui revient donc à corriger la popularité d'une page web par la prise en compte de son ancienneté. Les tableaux 1 et 2 fournissent, à titre d'illustration, une comparaison entre le top 10 renvoyé par Google (tableau 1) et le top 10 renvoyé si on classe les pages selon un rapport décroissant entre nombre de liens entrants réels et le nombre théorique (tableau 2).

Rang de la page sur google	Adresse url de la page web
1	www.veille.com/
2	www.veille.com/labo.htm
3	fr.groups.yahoo.com/group/CyberIES/
4	fr.groups.yahoo.com/group/intelligence-economique/
5	www.ege.eslsca.fr/
6	www.acrie.fr/
7	www.chez.com/ieco/
8	www.chez.com/ieco/Biblio.htm
9	www.geoscopie.com/acteurs/a532int.html
10	www.geoscopie.com/sources/internet/g083defesp.html

Tableau 1 : Classement renvoyé par Google pour la requête intelligence économique

classement sur google	url	date de référencement dans web.archive.org	nbre de liens entrants	nbre de liens entrants théoriques	rapport entre valeur observée et valeur théorique : clé de tri par valeur décroissante
52	www.strategic-road.com/intellig/ieconclass.htm	25/11/2002	162	3,481156842	46,82351511
12	c.asselin.free.fr/	05/10/2001	214	4,814324141	44,65839725
16	www.archimag.com/knowledge/	29/03/2002	83	3,481156842	24,12990963
136	www.ie-news.com/	17/05/2003	55	2,517165983	22,24724169
15	www.netpme.fr/intelligence_economique.html	02/06/2002	63	3,481156842	18,38469305

19	www.cybion.fr/	21/12/1996	418	24,35532955	17,20362679
123	perso.club-internet.fr/isaintel/	15/12/2000	110	6,658050179	16,67154753
1	www.veille.com/	06/12/1998	205	12,73416688	16,17695149
54	www.developpement-local.com/article.php3?id_article=182	15/01/2003	26	2,517165983	10,72634867
14	www.guerreco.com/	02/03/2001	42	4,814324141	8,931679451

Tableau 2 : Classement renvoyé requête intelligence économique en corrigeant l'analyse relationnelle par la prise en compte de l'âge de la page web

Sans porter de jugement sur ces deux classements respectifs, on observe que le classement est fort différent.

4. La pyramide des âges : un indicateur de veille sectorielle sur internet.

Lorsqu'on s'intéresse à un corpus de pages sur internet, il peut être intéressant de décrire ce corpus par un certain nombre d'indicateurs macroscopiques présentés dans un tableau de bord. Notre objectif est d'ajouter aux indicateurs traditionnels un indicateur temporel : la pyramide des âges d'un ensemble de site web. La pyramide des âges est traditionnellement un indicateur démographique. L'objectif de ce papier est, à la manière de ce qu'ont pu faire Douglis et al [3] et Pitkow et Pirolli. [10], de transposer au monde du web des indicateurs démographiques. La pyramide des âges d'un ensemble de pages web est un graphe qui représente la statistique de date de dépôt d'un nom de domaine en fonction de l'année de dépôt. Cet indicateur présente un intérêt en tant que tel mais il peut aussi être comparé entre plusieurs corpus.
Cet indicateur a fait l'objet d'une validation empirique sur 10 villes françaises réalisée en Mai 2004. [1]. Pour ce faire, nous avons utilisé le protocole suivant :

- On s'intéresse à 10 villes françaises (Dijon, Besançon, Quimper, Avignon, Aurillac, Belfort, Auxerre, Toulon, Annecy, Perpignan).
- Pour chacune de ces villes, l'expérience consiste à récupérer un échantillon de pages web permettant de caractériser la présence de ce territoire sur le web. Cet échantillon est constitué à partir du moteur de recherche Google. Pour chacune des 10 villes, on lance la requête *allintitle : nom de la ville*. Cette requête permet de récupérer l'adresse url des pages web dont le titre comporte le nom de la ville. On peut donc considérer que les réponses à cette requête renvoient plus que tout autre des pages caractérisant la présence de ce territoire sur internet.
- L'échantillonnage consiste à sélectionner les 200 premiers sites différents renvoyés par le moteur Google.
- Pour chacun de ces 200 premiers sites, on lance une requête whois grâce à une commande Linux ad hoc qui permet d'accéder aux bases de données des registers. Ces données sont ensuite parsées automatiquement à la recherche de la date de dépôt du nom de domaine.
- A partir des dates de dépôt des divers noms de domaine, il est possible de reconstituer la pyramide des âges du territoire. Cette pyramide des âges est biaisée par le fait que le moteur de recherche utilisé a tendance à valoriser les pages anciennes qui ont plus de chances que les autres d'avoir capitalisé un nombre significatif de liens entrants.

Ce travail a permis figure 5 de représenter la pyramide des âges de ces 10 territoires.

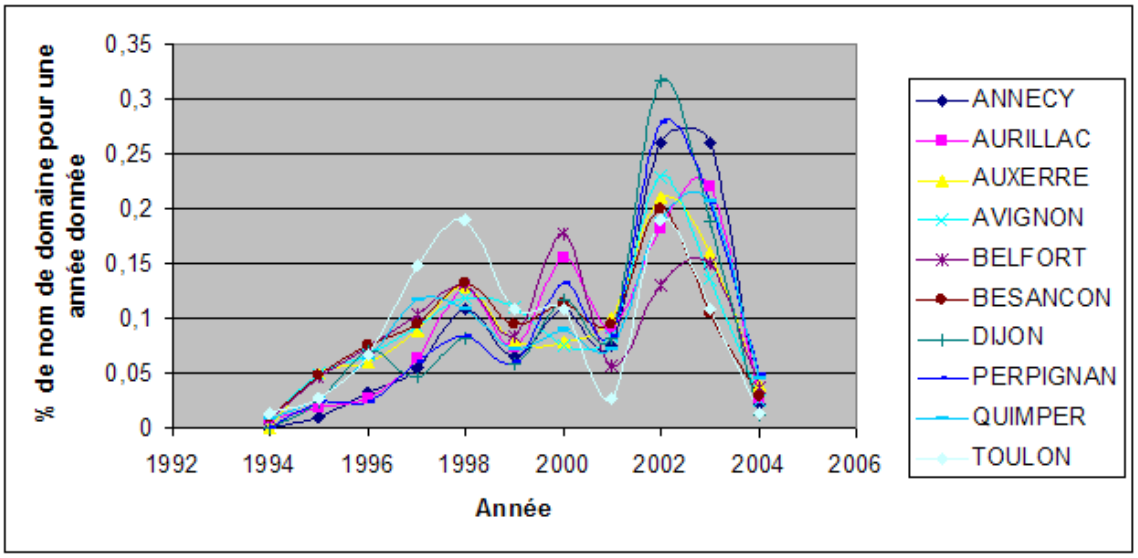

Figure 5 : pyramide des âges d'un corpus web

Les pyramides des âges sont représentées chacune par des cycles de deux ans. La superposition de ces 10 pyramides des âges est frappante. Difficile à interpréter, ces résultats soulèvent plusieurs pistes qui restent à développer :

- Ce caractère cyclique est il propre au domaine utilisé ou le retrouve-t-on dans d'autres domaines ?
- On a pu observer que les corpus de ces 10 villes n'étaient pas indépendants mais qu'il y avait des sites web que l'on retrouvait pour un grand nombre de villes. Ces sites expliquent ils à eux seuls la configuration du résultat ?

Une analyse plus fine peut être conduite en s'intéressant non plus aux sites web d'un ensemble de pages web d'un domaine mais aux pages web elles mêmes. On considère alors les dates de mise à disposition de ces pages sur internet en utilisant web.archive.org. C'est ce que nous avons réalisé figure 6 pour la requête intelligence économique. On observe des similitudes avec la courbe présentée Figure 5 bien que la source d'information soit différente.

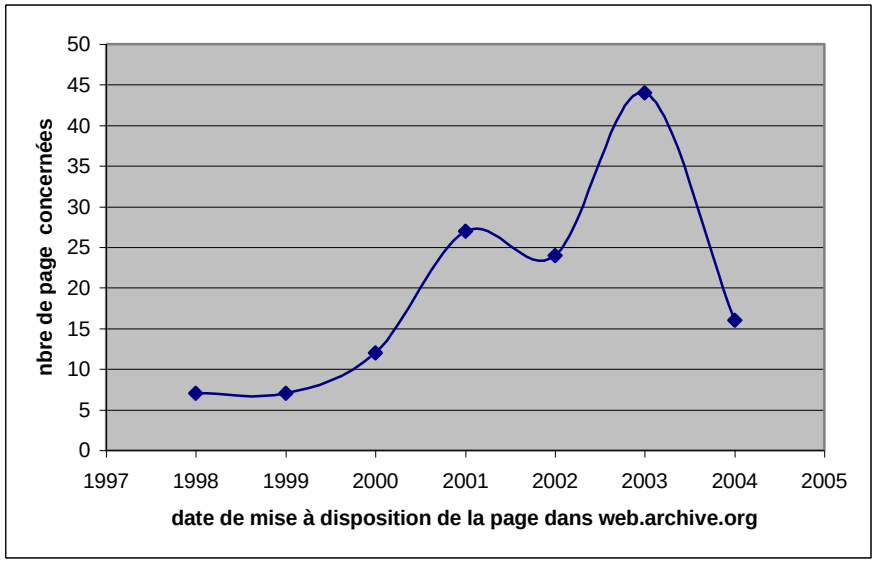

Figure 6 : pyramide des âges de la requête intelligence économique

5. Conclusion

Ce travail expérimental avait pour objectif à partir d'exemples concrets de montrer l'intérêt que peuvent représenter des indicateurs temporels pour l'analyse cybermétrique. Cette étude est un point de départ qui devrait déboucher sur des procédures permettant de systématiser l'analyse dynamique microscopique et macroscopique que nous avons mise en œuvre dans ce travail de façon encore semi automatique.
Ce travail de recherche souffre pourtant certaines limites :

- Il est tributaire des sources d'information qui ne sont pas toujours disponibles ni pertinentes
- Les principes présentés ont été mis en œuvre de façon semi automatiques ce qui rend difficile la duplication de l'expérience
- Il repose sur une base expérimentale légère qui ne permet pas de valider statistiquement les hypothèses qui ont été proposées

Toutefois ce document ouvre des pistes de recherches originales, un peu à la manière des travaux de Brewington et Cybenko [2] et propose une prise en compte du caractère dynamique du web.

6. Bibliographie

[1] BOUTIN E, Qualifier la présence d'une ville sur le web par des indicateurs cybermétriques dynamiques : une validation expérimentale sur 10 villes françaises. » Acte du colloque TIC et Territoires, Lille, Juin 2004

[2] BREWINGTON BE., CYBENKO G. How dynamic is the web? In Proceedings of the Ninth WWW Conference, Amsterdam, The Netherlands, 2000

[3] DOUGLIS F., FELDMANN A., KRISHNAMURTHY B. Rate of change and other metrics: a live study of the world wide web. In Proceedings of the USENIX Symposium on Internet Technologies and Systems, Monterey, 1997.

[4] FEISE J., Accessing the History of the Web: A Web Way-Back Machine". Open Hypermedia Systems and Structural Computing : 6th International Workshop, OHS-6, 2nd International Workshop, SC-2, San Antonio, Texas, USA, May 30 – June 4, 2000 p. 38-45

[5] FETTERLY D. MANASSE M., NAJORK M., WIENER JL. A large-scale study of the evolution of web pages. In Proceedings of the Twelfth WWW Conference, Budapest, Hungary, 2003.

[6] NTOULAS A., CHO J., OLSTON B. What's New on the Web? The Evolution of the Web from a Search Engine Perspective, WWW2004, May 17–22, 2004, New York, New York, USA, page 1

[7] PAGE L., BRIN S., MOTWANI R., WINOGRAD T. The PageRank citation ranking: Bringing order to the web. Technical report, Stanford Digital Library Technologies Project, 1998. Paper SIDL-WP-1999-0120 (version of 11/11/1999).

[8] The WebArchive Project, UCLA Computer Science, http://webarchive.cs.ucla.edu.

[9] LIM L., WANG M., PADMANABHAN S., VITTER JS., AGARWAL RC., Characterizing web document change. In Proceedings of the Second International Conference on Advances in Web-Age Information Management, pages 133–144. Springer-Verlag, 2001.

[10] PITKOW J., PIROLLI P., Life, death, and lawfulness on the electronic frontier. In Proceedings of the ACM Conference on Human Factors in Computing Systems, Atlanta, Georgia, 1997.