Qualifier la présence d'une ville sur le web par des indicateurs cybermétriques dynamiques : une expérimentation sur 10 villes françaises

Eric Boutin

QUALIFIER LA PRESENCE D'UNE VILLE SUR LE WEB PAR DES INDICATEURS CYBERMETRIQUES DYNAMIQUES : UNE VALIDATION EXPERIMENTALE SUR 10 VILLES FRANÇAISES.

Eric Boutin
Maître de Conférences
IUT de Toulon – Laboratoire le Pont
BP 132 83957 La Garde Cedex
boutin@univ-tln.fr

Résumé : Cet article s'intéresse à la datation d'une page web et à l'analyse de l'évolution de celle-ci depuis sa création. L'étude de ces variables temporelles débouche sur des indicateurs macroscopiques et microscopiques. Cet article fait l'objet d'une validation expérimentale dans le domaine de la veille territoriale à partir de l'étude de la présence sur le web de 10 villes françaises.

Summary : This paper is concerned with giving a date to a web page and analyzing the evolution of a web page since its creation. The study of these temporal variables creates macroscopic and microscopic indicators. This paper is validated in the area of territorial intelligence through the study of the presence on the web of 10 french cities..

Mots clés : intelligence territoriale, pyramide des ages, cybermétrie, mémoire du web

Qualifier la présence d'une ville sur le web par des indicateurs cybermétriques dynamiques : une validation expérimentale sur 10 villes françaises.

Dans cet article, on s'intéresse à la mesure de la présence de villes sur le web. Pour des raisons liées au caractère expérimental de la démarche, nous nous sommes limités à l'étude de la présence de 10 villes françaises. Nous proposons de mettre en œuvre des indicateurs cybermétriques dynamiques. Ces derniers ont pour objectif, non pas de qualifier la présence d'une ville sur le web à un instant donné mais de révéler quelles ont été les évolutions de cette présence au cours du temps. Dans un premier temps nous présenterons un état de l'art des indicateurs cybermétriques appliqués à la veille territoriale. Cette analyse permettra de positionner cette recherche et de faire ressortir son potentiel innovant. Ces indicateurs dynamiques seront ensuite déclinés autour de la dichotomie : indicateurs macroscopiques / indicateurs microscopiques.

1- ETAT DE L'ART DES INDICATEURS CYBERMETRIQUES APPLIQUES A LA VEILLE TERRITORIALE

Ce travail s'inscrit dans la problématique de la construction d'indicateurs visant à refléter la présence d'un territoire sur internet [Bertacchini, Boutin, 2003]. Cette problématique débouche sur des tableaux de bord permettant de décrire un territoire sous divers angles. Il existe donc selon Bertacchini (2003) plusieurs familles d'indicateurs. Chacun d'eux est caractérisé par son niveau d'analyse et par la technologie qu'il met en œuvre.
Deux niveaux d'analyse et trois technologies peuvent être utilisés pour caractériser la présence d'un territoire sur le web.

1.1 - Les deux niveaux d'analyse : macroscopique ou microscopique

Lorsqu'on effectue une analyse macroscopique sur un corpus web, on s'intéresse à un ensemble primaire de pages caractérisant le territoire. Ces pages sont bien souvent obtenues à partir d'un moteur de recherche sur internet. Ces pages sont ensuite qualifiées à partir de certains critères en recourant par exemple aux fonctionnalités avancées des outils de recherche. Voici à titre d'illustration quelques exemples d'analyses macroscopiques qui peuvent être effectuées :
Quel est le % de pages francophones sur le sujet
Quelle est la pyramide des âges des sites web parlant du sujet
Quel est le niveau d'interaction entre ces pages web
Quelle est la thématique de ces pages : pages institutionnelles, marchandes, personnelles…
Les indicateurs microscopiques vont s'intéresser à un nombre plus restreint de page web. Ces pages vont alors pouvoir faire l'objet d'une analyse plus qualitative. Voici quelques indicateurs qui peuvent être construits dans cette perspective :
Quel est le contenu de la page ?
Quel est le degré d'ouverture de la page vers d'autres pages ?
Quelle est la légitimité de la page parmi les pages du domaine ?
Quelle est la fraîcheur des informations contenues sur la page ?
Quelle est la dynamique d'évolution de cette page au cours du temps ?

1.2 - Les technologies sous jacentes : analyse de contenu, analyse relationnelle ou analyse dynamique

L'analyse de contenu conduit à extraire des informations du contenu textuel d'un document. Ce type d'analyse met en œuvre un processus d'analyse lexicale ou syntaxique à partir d'une information contenue dans une page web, l'adresse url d'une page web, le résumé contextuel d'un moteur de recherche
L'analyse relationnelle conduit à s'intéresser à l'interaction qui existe entre un ensemble de pages web et à caractériser ce niveau d'interaction par des indicateurs (de centralité, de densité…). L'analyse des réseaux sociaux présentée par Wasserman et Faust (1994) donne un cadre théorique à ce type de problématique.
L'analyse dynamique, qui nous préoccupe ici, conduit à s'intéresser à l'évolution d'une page web au cours du temps. Quand cette page web a-t-elle été créée ? De quelle manière cette page web a-t-elle évoluée au cours du temps ?

1.3 Croisement des technologies avec le niveau d'analyse

En recoupant l'information sur le niveau d'analyse et l'information sur la technologie sous jacente, on peut construire une matrice. Nous avons choisi de faire figurer dans chaque case de la matrice un exemple d'indicateur qui peut être défini.

	Analyse de contenu	Analyse relationnelle	Analyse dynamique
Analyse macroscopique	Quel est le % de pages francophones sur le sujet ?	Quelle est la densité des interactions entre un ensemble de pages web ?	Quelle est la pyramide des âges des pages web d'un sujet ?
Analyse microscopique	Quel est le contenu des thématiques abordées sur la page ?	Quelle est la légitimité de la page parmi celles du domaine ?	De quelle manière une page web a-t-elle évolué au cours du temps ?

L'objectif de ce travail est de montrer tout le potentiel de l'analyse dynamique dans le domaine microscopique et macroscopique. Nous avons choisi de privilégier cette analyse pour plusieurs raisons :
C'est une analyse très peu couverte par la littérature actuelle. La raison principale est que les fonctionnalités actuellement disponibles dans les outils de recherche rendent difficile la valorisation d'une telle analyse. La seule information temporelle disponible dans le contenu html d'une page web est la date de dernière mise à jour de la page web. La personne souhaitant faire une analyse dynamique d'une page web doit avoir un recul suffisant et stocker manuellement l'état de cette page à divers moment dans le temps. La mise à disposition de l'outil de recherche *web.archive.org* bouleverse les choses.
Nous avons eu l'occasion de nous pencher dans un passé récent sur les autres familles d'indicateurs et de présenter un indicateur dans le domaine de l'analyse dynamique macroscopique qui consiste à dresser la pyramide des âges des sites web traitant d'un sujet. Cette étude préalable nous a permis d'automatiser une démarche jusqu'alors manuelle et nous a conduit à poursuivre la réflexion en analysant aujourd'hui les choses sous un angle microscopique.
C'est une analyse qui possède un potentiel intéressant : Lorsqu'on interroge un moteur de recherche sur une thématique particulière, on obtient une photographie à un instant donné de l'état des ressources disponibles sur ce domaine. Or ce qui peut apparaître comme un ensemble figé est l'héritage d'une histoire qu'il s'agit de faire revivre. C'est cette dynamique qui nous intéresse ici.

2. LES INDICATEURS CYBERMETRIQUES DYNAMIQUES : analyse macroscopique et microscopique

2.1 La « pyramide des âges » de 10 villes françaises

Une ville sur internet est présente à travers plusieurs milliers de pages qui font référence à cette ville. Ces pages ont été créées à des dates différentes en fonction de critères qui sont endogènes ou exogènes au territoire. Il nous a semblé important de définir un indicateur, transposé du domaine démographique, qui corresponde à la notion de pyramide des âges des sites web d'un territoire. Notre objectif a été d'automatiser une démarche permettant de dresser la pyramide des âges d'un territoire et de comparer entre elles plusieurs pyramides des âges. Afin de présenter ce travail de façon didactique, nous présenterons d'abord le protocole utilisé, les limites inhérentes à la démarche, les résultats obtenus et les interprétations qui s'en dégagent.

Protocole de l'expérience :

On s'intéresse à 10 villes françaises (Dijon, Besancon, 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 bases de données comportent des informations caractérisant le nom de domaine déposé (nom du déposant, adresse et coordonnées du déposant, date de dépôt…)
L'encadré 1 fournit un exemple de résultat obtenu par le lancement d'une commande whois sur le site encyclopedia.com

```
<whois domain="encyclopedia.com">
[Requête en cours whois.internic.net]
[Redirigé vers whois.godaddy.com]
[Requête en cours whois.godaddy.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
    590 N. Gulph Rd
```

King of Prussia, Pennsylvania 19406
United States
+1.6109718840 Fax --

…
</whois>

Encadré 1 : exemple de données obtenues par une requête whois

Ces données sont ensuite parsées automatiquement à la recherche de la date de dépôt du nom de domaine. La tache est rendue malaisée par le fait que les données fournies par les registers ne sont pas structurées de façon homogène.
A partir des dates de dépôt des divers noms de domaine, il est possible de reconstituer la pyramide des âges du territoire

les limites de l'expérience :

Le protocole que nous avons réalisé présente deux limites principales :
On observe qu'il est très difficile d'extraire automatiquement la date de dépôt du nom de domaine de noms de domaines qui sont déposés en .fr pour des raisons liées à la structuration des données renvoyées par le register qui traite ces données. Souhaitant arriver à une automatisation de la démarche, nous avons donc fait le choix de ne pas prendre en compte les noms de domaines déposés en .fr ce qui est dérangeant lorsqu'on s'intéresse à la présence de villes françaises sur le web
L'information relative à la date de dépôt du nom de domaine n'a aucun sens pour des sites communautaires. En effet un site personnel déposé sur un espace communautaire n'aura pas obligatoirement pour date de création la date de création du site communautaire lui-même. Nous avons donc dû supprimer de l'analyse les sites personnels hébergés par des sites communautaires.

principaux résultats obtenus :

Nous allons présenter successivement la pyramide des âges d'une ville (Perpignan) puis la pyramide des âges agrégée des 10 villes que nous avons analysées.
La figure 1 présente le résultat de la pyramide des âges de la ville de Perpignan.

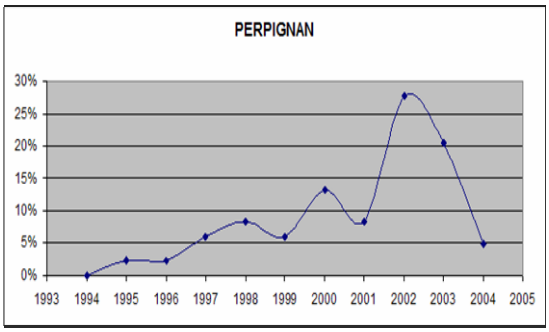


Figure 1 : pyramide des âges de la ville de Perpignan

Ce graphe se lit de la façon suivante :
En abscisse, on trouve les dates de dépôt des noms de domaine
En ordonnées, on mentionne le pourcentage de noms de domaine déposés pour une date donnée
Le graphe représentant la pyramide des âges de Perpignan fait apparaître une structure cyclique qui tend à se répéter tous les deux ans. On observe en effet que les dépôts des noms de domaines ont connu des pics en 1998, 2000, 2002.

Lorsqu'on représente sur un même graphe les pyramides des âges des 10 villes françaises, on obtient le graphe de la figure 2.
Nous avons été frappés par la ressemblance forte existante entre ces pyramides des âges. Rien ne nous laissait présager une superposition si claire des 10 pyramides des âges. En effet, les villes analysées ont toutes une expérience de la présence sur le web qui est unique et il est curieux que l'on retrouve de tels invariants à une telle échelle.
Il est difficile de passer du stade de la description factuelle au stage de l'interprétation. Nous en sommes réduits à livrer quelques pistes auxquelles une analyse plus approfondie nous permettra de répondre :
Ces cycles bi annuels sont ils généraux dans toute analyse web ou les observe-t-on uniquement lorsqu'on s'intéresse à des territoires ?
On a pu observer que les données de ces différentes villes ne sont pas indépendantes. Il existe en effet des sites qui sont présents pour plusieurs de ces villes. Ces sites expliquent-ils à eux seuls la similitude entre ces pyramides des âges ?

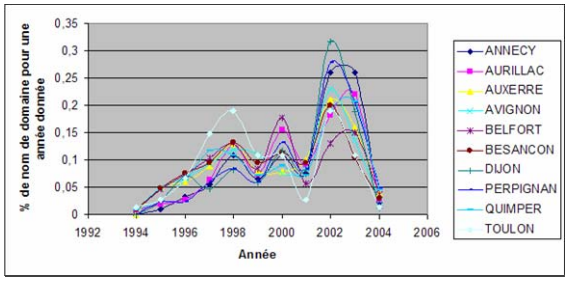


Figure 2 : pyramide des âges de 10 villes françaises

2.2 Analyse microscopique de la dynamique du site officiel de 10 villes françaises

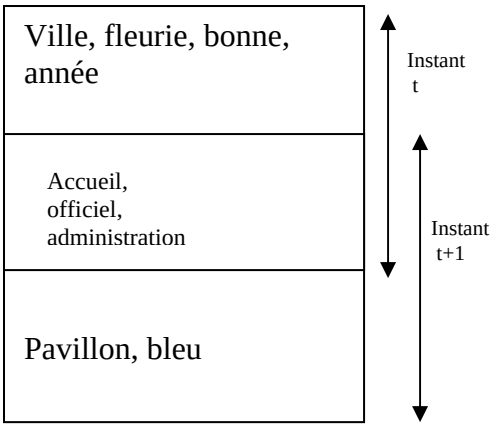
Notre analyse porte sur le suivi depuis 1998 de la page d'accueil du site web officiel de 10 villes qui sont des chefs lieux de départements français. L'identification du site officiel d'une ville est un problème simple qui peut être résolu en utilisant les fonctions élémentaires d'un moteur de recherche. Une fois ce travail effectué, nous avons donc une liste de 10 pages web. Le travail consiste alors à récupérer le contenu de ces pages web à divers instants dans le temps et à analyser cette évolution par des indicateurs. Dans la logique du paragraphe précédant, nous présenterons d'abord le protocole utilisé, les limites inhérentes à la démarche, les résultats obtenus et les interprétations qui s'en dégagent

Protocole de l'expérience :

Nous considérons au départ l'adresse web des pages d'accueil des sites web officiels des 10 villes que nous avons choisies.
La récupération de l'état de ces pages à différentes dates est rendue possible par l'utilisation du moteur *web.archive.org*. Ce moteur de recherche est le fruit d'un travail d'archivage du web afin, selon Feise (2000) 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 intervalle régulier. 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 Figure 3
donne l'exemple du résultat obtenu pour le site de la mairie d'Avignon.(mairie-avigon.fr)

Figure 3 : capture écran de web.archive.org sur la requête mairie-avignon.fr

Plusieurs observations peuvent être faites à partir de ce tableau :
Lorsqu'on clique sur une date on visualise l'état de la page web à cet instant donné.
Le nombre de captures du web par année a été différent au fil du temps : avant 2000, les pages étaient rescannées en moyenne 2 fois par an. Depuis 2002, les pages sont scannées parfois plusieurs fois par mois.
Les passages du robot ne s'effectuent pas au même moment sur toutes les pages web pour des raisons logistiques. Toutefois, on observe que le nombre de fois où le robot passe par an est globalement constant pour toutes les pages.
Cet outil dispose d'une information relative à l'existence d'une modification dans la page entre deux observations. Ces modifications sont représentées par une astérisque en face de la date ou une modification s'est produite. Précisons que l'existence d'une astérisque en face d'une date ne signifie pas que la page a été modifiée ce jour là mais que ce jour là le moteur a récupéré une page dont le contenu était différent du contenu observé l'instant précédent.
Cette source d'information a servi de point de départ à l'analyse que nous avons conduite. Pour procéder à cette analyse, nous avons cherché à quantifier le pourcentage de modification intervenu dans une page web entre deux périodes de temps. Pour ce faire, nous avons privilégié le contenu textuel de la page web entre les dates de modification prises deux à deux. L'analyse de l'évolution du texte est alors caractérisée par un indicateur.
Considérons l'exemple ci-dessous pour expliciter cet indicateur: Soit une page simple définie par les mots suivants à l'instant t :
Accueil
Officiel
Administrations
ville
Fleurie
Bonne
année
A l'instant t+1, la page d'accueil de cette ville comporte le texte suivant :
Accueil
Officiel
Administrations
Pavillon
bleu
L'évolution du contenu textuel de cette page peut être représenté par un diagramme qui permet de visualiser le texte commun à ces deux pages et le texte spécifique.



Nous avons défini un indicateur qui s'apparente au % de modification introduit dans une page entre deux espaces de temps. Cet indicateur qui vaut 1 lorsqu'aucun des mots clés présents à l'instant t ne se retrouve à l'instant t+1 et 0 si les mots clés de la page se retrouvent à l'identique entre deux espaces de temps successifs.
Nous avons retenu l'indicateur de Bray et Curtis. Cette indicateur est complémentaire à 1 de celui de Czekanowski.
L'indicateur de Czekanowski est le rapport entre le double du nombre de mot clé commun entre les deux versions d'une page et la somme du nombre de mots clés apparaissant dans chacune des versions de la page. Dans l'exemple considéré, on obtient 3*2/(7+5) soit un indicateur de Czékowski de 6/12. L'indicateur de Bray et Curtis est donc de 6/12 soit 0.5.

Limites de l'analyse :

Il existe une limitation forte associée au fait que le % de modification d'une page entre deux espaces de temps n'est apprécié que par le texte de cette page. La composante graphique n'est donc nullement prise en considération. Cette composante graphique est plus ou moins forte pour les villes étudiées et comporte parfois du texte ou des menus qu'il n'est donc pas possible d'analyser en tant que tel.
Nous avons raisonné pour chacune des 10 villes sur la page web qui s'affiche lorsqu'on arrive sur le site officiel. En fonction de la structure du site web, cette page n'est pas forcément riche de contenu. Il s'agit parfois d'une image très pauvre en texte.
Ces deux raisons font qu'il ne faut pas prendre ces analyses au pied de la lettre en réduisant la dynamique d'une ville sur le web à l'analyse réductrice que nous avons menée.

principaux résultats obtenus :

Le graphe présenté figure 4 illustre graphiquement l'évolution des 10 pages de garde des sites web depuis le premier janvier 1999.
Ce graphe s'interprète selon la grille de lecture suivante :
L'abscisse précise le temps.
En ordonnées, on mentionne l'indicateur cumulé de Bray & Curtis pour la ville considérée. Si on s'intéresse à une ville donnée, entre deux paliers, on observe un écart maximum de 1. Un tel écart signifie que la page n'a plus du tout les mêmes mots clés qu'à l'instant d'avant.

Ce graphe fait ressortir le rôle de certaines villes qui utilisent leur page de garde de site web pour communiquer des informations régulièrement mises à jour. La ville d'Avignon illustre particulièrement cette approche. La valeur de 14 associée à l'ordonnée signifie que les modifications textuelles affectées à cette page correspondent à l'équivalent d'un renouvellement du texte 14 fois.

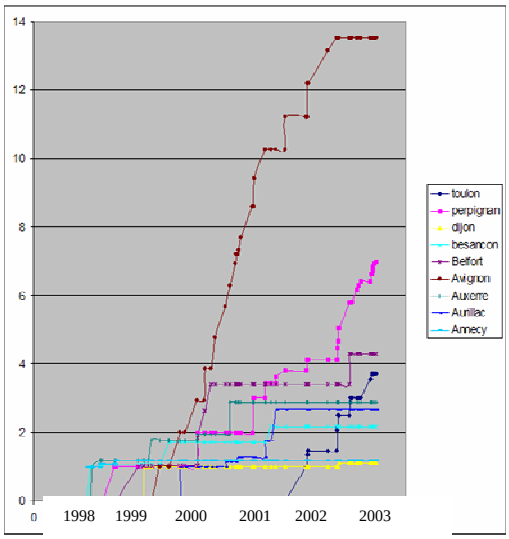


Figure 4 : évolution de la dynamique de la page de garde des sites web de 9 villes françaises

Dans cet article nous proposons quelques pistes d'analyse dynamique débouchant sur des indicateurs permettant de caractériser la présence de pages web de villes françaises. Cette étude est un point de départ qui a pour objectif de révéler la richesse de l'analyse dynamique du web et devrait déboucher sur des procédures permettant de systématiser l'analyse dynamique microscopique que nous avons mise en œuvre dans ce travail de façon encore semi automatique.

BIBLIOGRAPHIE

Feise J., (2000), “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*

Bertacchini Y., Boutin E. (2003), « Une lecture possible du territoire Sophipolitain : l'observation des représentations virtuelles d'une technopole » , *ISDM n°7 - Avril 2003*

Bertacchini Y. (2003) « Territoire physique/territoire virtuel : quelle cohabitation ? » *ISDM n°9 - Juillet 2003*

Wasserman, Faust, (1994) “Social networks analysis : methods and applications”, *Cambridge University Press*