Un outil de mesure de l'audience d'un site Internet : l'analyse réseau

Jean-Marc Ferrandi, Eric Boutin

UN OUTIL DE MESURE DE L'AUDIENCE D'UN SITE INTERNET : L'ANALYSE RESEAU

Jean-Marc FERRANDI *

Eric BOUTIN **

* Maître de Conférences, Université de Bourgogne, IUT de Dijon, Route des Plaines de l'Yonne, BP 58, 89010 AUXERRE, doc.iut89@demeter.fr.

** Maître de Conférences, Laboratoire Lepont, Université de Toulon et du Var, IUT de Toulon, BP 132, 83957 La Garde Cedex, boutin@univ-tln.fr

UN OUTIL DE MESURE DE L'AUDIENCE D'UN SITE INTERNET : L'ANALYSE RESEAU.

Résumé: La commercialisation sur Internet bouleverse les techniques commerciales traditionnelles. Pour optimiser l'audience de son site, voire ses ventes sur ce canal, l'entreprise a besoin d'informations statistiques pertinentes. Le but de notre recherche est de renouveler l'approche des analyseurs de fichiers *.Log*, outils disponibles actuellement, en montrant les apports de l'analyse réseau.

USING NETWORK ANALYSIS AS A TOOL FOR MEASURING AN INTERNET SITE AUDIENCE.

Internet sales deeply change the traditional commercial techniques. To optimize the audience of its site, and more its sales on this media, an enterprise now needs relevant statistical information. Our research intends to renew the Log files analyzers, currently available tools, by showing the contributions of the network analysis.

Sur Internet, une entreprise, une université ou une organisation a la possibilité d'offrir à la communauté une information sous la forme de pages au format html liées les unes aux autres par des liens hypertextes. L'objectif poursuivi peut être de constituer une vitrine de ses activités, d'être présent sur un nouveau canal de communication, de répondre à une stratégie d'image ou de prestige, d'appuyer une stratégie de commercialisation de ses produits.

Le processus de communication sur le Web diffère de la démarche classique. Contrairement aux médias traditionnels, Internet permet aux clients de prendre l'initiative de la communication. En outre, à la différence de la communication télévisuelle, le visiteur est actif. Sa recherche part d'une démarche volontaire : demandeur d'informations sur un thème particulier, la durée de sa visite dépendra de la qualité de la réponse qui lui sera fournie. Aussi est-il nécessaire d'appliquer à la confrontation entre l'offre et la demande d'informations une des règles élémentaires du marketing : l'offre doit s'adapter à la demande.

Si l'on s'intéresse à la vente sur Internet, la navigation sur un site marchand s'apparente à la visite d'un magasin traditionnel. Dans ce dernier, l'entreprise a besoin de connaître les produits les plus demandés, d'améliorer leur agencement pour optimiser ses ventes. Sur un site, elle a besoin de savoir quelles sont les pages les plus consultées, les liens les plus utilisés. Une fois cette analyse réalisée, elle peut aménager son site ou son magasin virtuel en fonction des attentes de ses clients.

La mesure de l'audience d'un serveur Web répond aussi à un autre souci commercial. Certains sites vivent des bannières publicitaires qu'ils offrent (Onnein-Bonnefoy, 1997). Il est alors important de pouvoir justifier d'un prix à payer par l'annonceur en fonction de la page sur laquelle sa bannière se trouve et de savoir où ces panneaux doivent être positionnés pour réaliser une optimisation commerciale du site.

Il est donc essentiel pour l'entreprise de disposer de capteurs dans son environnement qui lui permettent de recueillir des informations relatives à la visite de son site, et d'être capable de traiter ces informations de manière à les rendre intelligibles.

Nous nous intéresserons ici à une source d'information incomplète mais toujours disponible : le fichier *.Log*[1] qui enregistre les connexions des différents utilisateurs. Son analyse, sur une période donnée, permet de mesurer l'audience d'un site en dégageant des invariants et en répondant aux questions suivantes :

- Quels sont les points de passage obligés du visiteur lorsqu'il se connecte à un site ?
- Quel est le parcours d'un visiteur type ?
- Lors d'une visite, comment s'articulent entre elles les différentes thématiques du site ?

- Y a-t-il sur le site des pages obsolètes visualisées un nombre de fois non significatif ?
- En terme de cheminement, quelles sont les pages qui accueillent les visiteurs ? Sur quelles pages quittent-ils le serveur ?

Ces éléments statistiques fournissent de précieuses indications sur le mode d'utilisation du site et permettent à l'entreprise de l'adapter aux besoins. Le Centre d'Etude des Supports de Publicité retient d'ailleurs quatre indicateurs d'audience d'un site Web : le nombre de pages vues par chaque visiteur, le nombre de visites sur une période donnée, l'origine géographique des connexions et la durée de consultation par visite.

Notre recherche a pour objectif de renouveler l'approche des analyseurs de *.Log* utilisés dans le commerce en montrant les apports de l'analyse réseau. Cette nouvelle démarche, utilisée dans d'autres cadres de recherche en marketing (Iacobucci, 1996 ; Boutin, Ferrandi, Valette-Florence, 1997) et en sociologie (Degenne et Forsé, 1994 ; Wasserman et Faust, 1994), est présentée en s'appuyant sur l'audit du serveur du laboratoire du CRRM, Centre de Recherche Rétrospective de Marseille, réalisé en décembre 1996. L'étude porte sur l'analyse des 2869 connexions enregistrées. Celles-ci correspondent à la consultation de 10 259 pages.

Ce serveur expose les axes de recherche du laboratoire (présentés en annexe 1) et permet à des étudiants de troisième cycle d'héberger leurs pages html. Les thèmes de ces pages sont libres et ne touchent pas forcément au domaine de la recherche.

Il aurait été pédagogique de distinguer les étapes successives de collecte, de traitement et d'analyse des données. Toutefois, une telle démarche est impossible à respecter. En effet, l'analyse réseau regroupe de nombreuses microanalyses qui se complètent pour donner une vue d'ensemble du phénomène à analyser. Un même phénomène complexe peut se présenter sous plusieurs facettes, chacune renvoyant à une partie de la réalité. Procéder à un découpage séquentiel nous aurait conduit à une grande confusion dans la mesure où il aurait fallu présenter toutes ces microanalyses, puis tous les réseaux qui en sont issus. Nous avons préféré présenter chacune de ces microanalyses en montrant en quoi elles contribuent à apporter un élément de réponse à la question posée. Seule la partie collecte des données, qui est commune à l'ensemble des microanalyses réseau, fera l'objet d'une présentation générale. Les traitements et l'exploitation des réseaux résultants seront présentés dans un second temps. Nous procéderons enfin à l'évaluation globale de la méthode et montrerons ses apports.

I. LE FICHIER .LOG ET SON TRAITEMENT.

Nous exposerons les différentes formes que peuvent revêtir les fichiers *.Log*, les limites qui leur sont inhérentes. L'évaluation des différents logiciels de traitement de ces fichiers nous permettra de dégager l'opportunité de l'analyse réseau.

1. Les fichiers .Log.

Les fichiers *.Log*, aussi appelés fichiers traces, enregistrent sur un fichier texte les différentes actions effectuées par les visiteurs d'un serveur Web. Leur analyse permet de mesurer l'audience d'un site. Ces fichiers sont de quatre types :

- Les Log de transfert enregistrent les requêtes reçues par le serveur. Ils peuvent se présenter sous trois formats principaux : Common Log Format[2], format étudié dans le cadre de cette recherche, Extended Common Log Format et Havest.
- Les Log d'erreur gardent la trace des erreurs survenues lors du téléchargement d'une page par le visiteur.
- Les Log référentiels indiquent d'où vient l'utilisateur quand il se connecte sur le site.
- Les Log d'agent renseignent sur le type de navigateur (Nescape, Explorer,…) utilisé par les visiteurs du site.

Pour illustrer le caractère massif de l'information collectée, nous allons représenter trois des 2869 connexions. Le fichier résultant, figure 1, se présente comme une succession de lignes ou hits. Ces trois utilisateurs ont respectivement visualisé 5, 2 et 1 pages du serveur du Crrm.

```
194.51.254.3 - - [01/Dec/1996:01:36:46 -0100] "GET /cgi-bin/Count.cgi?tr=N&dd=C|df=polar.dat HTTP/1.0" 200 907
194.51.254.3 - - [01/Dec/1996:01:37:55 -0100] "GET /entertainment/polar/polarweb/pwtete.htm HTTP/1.0" 200 2440
194.51.254.3 - - [01/Dec/1996:01:38:28 -0100] "GET /entertainment/polar/polarweb/album.htm HTTP/1.0" 200 2344
194.51.254.3 - - [01/Dec/1996:01:42:09 -0100] "GET /entertainment/polar/polarweb/pwcritik.htm HTTP/1.0" 200 9407
194.51.254.3 - - [01/Dec/1996:01:45:28 -0100] "GET /entertainment/polar/polarweb/pwlinks.htm HTTP/1.0" 200 14973

202.131.0.29 - - [01/Dec/1996:12:12:12 -0100] "GET /vl/vlis.html HTTP/1.0" 200 1876
202.131.0.29 - - [01/Dec/1996:12:12:20 -0100] "GET /vl/metricts.html HTTP/1.0" 200 9841

crrm.univ-mrs.fr - - [01/Dec/1996:12:12:12 -0100] "GET /vl/vlis.html HTTP/1.0" 200 1876
```

Une Connexion

Une page visualisée

Figure 1 : Exemple de fichier .Log.

Au départ, le fichier *.Log* ne se présente sous cette forme que si un seul utilisateur peut se connecter à la fois sur le site analysé. En général, les différentes lignes ne sont pas classées par visiteur mais dans l'ordre de leur arrivée sur le serveur. Il est alors nécessaire de les trier en fonction de la date de connexion si on veut retrouver les données de la figure 1.

2. Les limites de l'information contenue dans les fichiers .Log et les solutions apportées.

L'information contenue dans les fichiers *.Log* doit être interprétée avec précaution en raison du mode d'identification de l'utilisateur et du mode de construction du fichier.

En effet, chaque utilisateur n'est pas identifié de façon univoque dans le fichier. Si deux visiteurs ayant le même nom de serveur se connectent au même moment, leur distinction sera impossible, ce qui introduit un risque de confusion potentielle. De plus, l'adresse du serveur ne permettra jamais de connaître le nom de l'utilisateur, ni son adresse électronique.

En outre, le fichier ne garde pas une trace fidèle des différentes pages visualisées par le visiteur. Lorsqu'un utilisateur souhaite visualiser une page qu'il a déjà fait apparaître, sa requête n'est pas toujours répercutée sur le site principal mais chargée à partir de la mémoire cache du navigateur ou de celle du proxy[3]. La retranscription de l'intégralité du cheminement d'un utilisateur sur un site donné est de ce fait impossible.

Le fichier *.Log* enregistre toute nouvelle page visualisée par le client du site. Cet enregistrement de l'information a deux incidences. L'enchaînement de deux pages dans le fichier ne signifie pas toujours que ces pages sont liées l'une à l'autre par un lien hypertexte direct. De plus, le différentiel de temps entre deux lignes, lors d'une même connexion, ne doit pas s'appréhender uniquement comme le temps passé sur une page donnée, mais comme le temps passé avant de visualiser une nouvelle page.

On peut gérer de deux façons le problème de la non-retranscription par les fichiers *.Log* de l'intégralité des hits visualisés par le visiteur.

La première consiste à introduire sur chaque page une petite image dynamique qui va forcer le serveur à se reconnecter sur le site pour rafraîchir l'image. Tout nouvel affichage débouche alors sur l'inscription de la page affichée sur le fichier *.Log*. On obtient ainsi un fichier fidèle sauf si le visiteur ne souhaite pas visualiser les images. Toutefois, appliquer cette méthode conduirait à alourdir la navigation sur Internet et à décourager le visiteur.

La seconde revient à reconstituer, à partir du fichier *.Log*, la stratégie suivie par l'utilisateur du site. Sur la base de l'hypothèse que le visiteur utilise le plus court chemin, on va considérer que, lorsqu'un lien entre deux pages ne correspond pas à un lien réel, on recherchera le plus court chemin permettant de passer de l'un à l'autre. Ce chemin, en théorie des graphes, est appelé géodésique. Sa détermination suppose que l'analyste ait à côté du fichier *.Log*, pour la période prise en compte, une information relative à tous les chemins géodésiques entre chaque paire de pages. Il faudrait, pour ce faire, considérer une période de temps où l'architecture du

site n'a pas été modifiée. Nous n'avons pas pris en compte ce type d'analyse dans cette recherche.

3. Le traitement des fichiers .Log.

Dans un premier temps, nous présenterons les résultats obtenus en utilisant les analyseurs de Log proposés dans le commerce pour dégager l'opportunité de l'analyse réseau. Cette analyse sera ensuite approfondie.

A/. Les analyseurs de fichiers .Log du marché.

Les analyseurs présents dans le commerce se situent en aval du fichier *.Log* et restituent une information de synthèse souvent sous la forme de tableaux statistiques. Nous exposerons successivement trois familles d'analyseurs[4] en fonction du degré de complexité croissante des analyses qu'ils proposent.

Analog 2.11[5] (1998) est un logiciel de type boîte noire. Il génère automatiquement un fichier de synthèse type, présentant les résultats sous la forme de tableaux statistiques à une dimension, semblables à des tris à plat. Ces tableaux décortiquent l'information relative à la date de connexion (ventilation par mois, par jour, par heure), à l'adresse du visiteur et aux pages visualisées par chaque visiteur.

Webtracker[6] (1998) présente trois différences fondamentales par rapport à Analog. Il illustre une famille de type boîte à outils. Les tableaux créés sont de type "tris à plat" et "tris croisés". Ce logiciel permet de croiser l'information contenue dans la variable temporelle de son choix (date, semaine, mois, année, heure, jour de la semaine, jour du mois, semaine ou mois de l'année) avec une autre variable (adresse, pays de provenance, page visualisée, nombre d'octets transférés). Enfin, chaque résultat peut être exprimé selon l'occurrence dans le corpus ou selon le nombre d'octets télédéchargés.

Hitlist[7] (1998) intègre une notion inexistante dans les deux familles précédentes : le concept de session[8]. Dans les familles précédentes, l'audience du site était appréhendée à travers le nombre de requêtes. Désormais, elle est également saisie à travers le concept de visite[9]. Cependant, dans le fichier *.Log*, les visites ne sont pas identifiées en tant que telles. Il faut reconstituer, sur la base de la succession des pages enregistrées dans le fichier, les pages visuali-

sées par chaque utilisateur. Dans Hitlist, on suppose que lorsqu'un visiteur attend plus de quinze minutes sur une page avant d'en lancer une autre, il a entamé une nouvelle session.
Le type de tableaux que propose ce logiciel fournit une première idée de l'activité associée à un site et aux différentes pages le constituant. Toutefois, cet outil ne va pas suffisamment loin dans l'exploitation du concept de visite : il se contente d'exploiter la page d'arrivée des visiteurs sur le site et la page de clôture sans véritablement s'intéresser au parcours des visiteurs.

B/ Un nouveau mode de traitement des fichiers .Log : l'analyse réseau.

Par rapport à ces trois outils, l'approche réseau apporte un supplément d'informations. Lorsqu'un client se connecte sur un site, les pages qu'il visualise sont porteuses de sens. Mais on peut aussi s'intéresser à l'ordre de visualisation de ces pages. Cet ordre prend en compte les liens retenus par le client. Les analyseurs commerciaux n'envisagent pas la dimension séquentielle de la consultation : ils présentent des informations indépendantes les unes des autres. Au contraire, l'analyse réseau enrichit cette information statique par le sens qui est donné aux liens, reconstituant ainsi la démarche de l'utilisateur.
La consultation d'une page par un grand nombre de clients peut être due à deux éléments en interaction : la qualité intrinsèque de la page et/ou sa position par rapport aux autres pages. L'analyse statistique classique ne permettra pas de juger de ce second critère. Au contraire, une représentation sous forme de réseau permettra de visualiser cette page dans son contexte et de la caractériser par un certain niveau de centralité.
Pour pouvoir mettre en œuvre l'analyse réseau, un traitement préalable des données est nécessaire. Son objectif est de structurer l'information en un format qui autorise l'exploitation des données par les outils de traitement automatique Dataview (Rostaing, 1993) et le logiciel d'analyse de réseaux Matrisme (Boutin, 1999).
La grande différence entre le fichier initial et le fichier formaté est que celui-ci cherche à reconstituer, à partir d'un ensemble de hits positionnés chacun sur une ligne, une connexion réalisée par un utilisateur sur le site. Comme un même visiteur peut se connecter à plusieurs reprises sur le site, si on souhaite isoler chaque séquence de travail, il faut considérer qu'après un certain laps de temps, une nouvelle connexion a commencé. Nous avons estimé qu'un arrêt de plus de dix minutes correspondait à une nouvelle session de travail. Cette variable est paramétrable.

Le traitement manuel des milliers de lignes du fichier *.Log* étant impossible et ces fichiers présentant une structure homogène, le formatage des données a été automatisé en utilisant une routine informatique (Webmap). Celle-ci permet, à partir des trois indicateurs de base fournis lors d'une connexion (nom du serveur du client, date de la connexion, nom de la page visualisée), de créer par combinaison cinq indicateurs supplémentaires.

Un exemple de la structure de l'information formatée à partir des cinq premières lignes du fichier de la figure 1 est exposé dans le tableau 1 qui présente une synthèse du vocabulaire utilisé. Sept champs renseignés par une ou plusieurs modalités sont définis. Cette information est plus riche que l'information brute disponible initialement : cinq informations supplémentaires apparaissent.

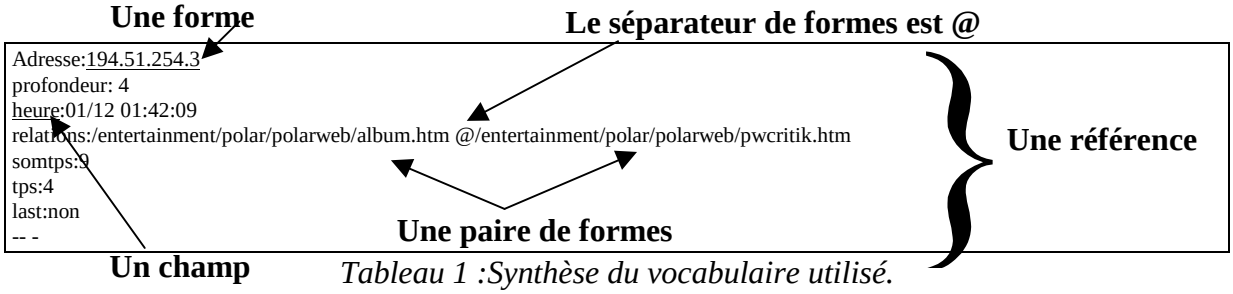

Tableau 1 :Synthèse du vocabulaire utilisé.

1. Le niveau de profondeur : il correspond au nombre de pages successives visualisées par le visiteur avant de faire apparaître la page active. Ainsi, une profondeur de 1 correspond à la page par laquelle l'utilisateur arrive sur le serveur. Plus la profondeur est élevée, plus le visiteur a navigué sur le site.

2. Les deux dernières pages visualisées : Lorsque l'utilisateur change de page, la rubrique «relations» indique les deux dernières pages regardées (la paire de formes). Cette rubrique dynamise l'information statique contenue dans le fichier initial. Pour un niveau de profondeur de 1, le champ relation est renseigné par la mention "ND" (Non Défini) car le visiteur vient de se connecter.

3. Le temps passé par un client entre deux pages : Pour une connexion donnée, il est déterminé en faisant la différence entre les dates prises deux à deux. Cette information contenue sous la rubrique "tps" est exprimée en minutes. Plus sa valeur est élevée :

- plus le temps passé sur la page elle-même est grand. Le temps passé sur une page dépend de l'intérêt que présente cette page pour l'utilisateur.

et/ou

- plus le temps de transfert entre cette page et la suivante est élevé. Celui-ci est une fonction croissante du nombre d'images que la page doit charger, du degré de saturation du

réseau à l'heure de la connexion, du nombre de pages intermédiaires à visualiser avant d'arriver sur une nouvelle page.

Il peut donc difficilement être interprété comme un indicateur de pertinence d'une page.

4. Le temps total passé par le client sur le site : il est obtenu en agrégeant le temps passé sur chaque page. Cette information contenue dans le champ «somtps» est exprimée en minutes.

5. La dernière page visualisée par un client : on l'identifie lorsque le champ "last" est oui.

Ces quelques indicateurs permettent de répondre aux questions suivantes: Quel est le point d'entrée sur le site ? Quel est le point de rupture de la connexion ? Quelles sont les pages les plus visitées ? A quelles thématiques correspondent-elles ? Quel est le temps moyen passé par un visiteur ? Quel est le lien entre chacun des thèmes ou chacune des pages visualisées par le visiteur ?

Il est possible de définir d'autres indicateurs en fonction des objectifs de l'étude : jour auquel a lieu la connexion, pays d'origine du client, lorsque cette information est disponible. Toutefois, nous n'avons pas retenu ici ces paramètres.

Le processus de formatage des données s'accompagne de l'élimination de deux types de données non pertinentes pour l'analyse. Les premières correspondent aux connexions établies sur le site en interne. Dans la mesure où l'objectif est de faire l'audit d'un serveur, il paraît plus judicieux d'analyser de façon disjointe ces connexions de celles effectuées à partir d'autres serveurs. C'est la raison pour laquelle les premières ont été retirées. D'autre part, lorsqu'une page se charge, chacune des images qui lui est associée fait l'objet d'une ligne supplémentaire dans le fichier. Si le but est de suivre les actions effectuées par le visiteur, la prise en compte de ces chargements automatiques est inutile. C'est pourquoi les pages, où figurait une extension ".gif" ou ".jpg" signalant la présence d'une image, ont été supprimées.

En passant d'une information brute à une information formatée, nous avons accompli un premier travail de création de valeur ajoutée.

II LA MESURE DE L'AUDIENCE DU SITE DU CRRM.

Notre approche repose sur la construction d'un graphe appelé réseau. Les sommets de ce graphe sont les différentes pages du site du CRRM. Un arc entre deux sommets signifie qu'un visiteur au moins est passé d'une page à l'autre. L'interprétation des réseaux peut se faire de manière très intuitive par l'observation visuelle du graphe, mais aussi en s'aidant d'indicateurs

de synthèse qui permettent de rationaliser l'analyse en extrayant un certain nombre de sommets aux propriétés particulières.
Nous considérerons successivement deux angles complémentaires du problème. Dans un premier temps, nous raisonnerons sur le réseau global, qui retranscrit toutes les connexions établies, quelles que soient leur durée ou leur profondeur. Ensuite, nous améliorerons notre compréhension du comportement des visiteurs en examinant quelques réseaux particuliers.

1. L'analyse du réseau global.

Le réseau qui visualise le parcours des 2869 visiteurs du site sur la période considérée renvoie à un graphe parfaitement inextricable : le réseau, graphe fidèle à la réalité, ne fait que retranscrire le réel avec le moins de déformation possible. Lorsque la réalité est complexe, par corollaire, le réseau l'est aussi.
Plusieurs analyses utilisant la technique du filtrage peuvent être menées pour dégager de ce réseau des informations pertinentes. Filtrer va consister, selon le cas, à supprimer du réseau global certains sommets ou certains liens ou les deux à la fois. Nous avons identifié trois types de filtrages possibles : le filtrage des paires, celui des formes et celui des connectivités.
Les filtres peuvent être mis en œuvre de manière manuelle ou automatique. Dans un filtrage manuel, la frontière entre les pages retenues et celles qui ne le sont pas est choisie par l'analyste en fonction de son expérience. Lors d'un filtre automatique, le découpage de l'information est proposé directement par le logiciel Matrisme. Toutefois, la détermination automatique des filtres n'est possible que si le graphe positionnant les éléments du champ considéré d'après leur fréquence décroissante se traduit par une représentation de type zipfienne (Zipf, 1949) illustrée figure 2.

Tague et Nicholls (1987) définissent la courbe zipfienne par la fonction $g_x = \frac{a}{x^b}$ où g_x représente le nombre de modalités apparaissant exactement x fois, a le nombre d'éléments apparaissant une seule fois et b la dispersion des fréquences des modalités. Le nombre de modalités correspondant à une fréquence d'apparition donnée est donc inversement proportionnel à cette fréquence.
Cette courbe traduit le fait qu'il existe sur le site du CRRM un petit nombre de pages qui sont très fortement visitées, un grand nombre de pages qui sont visualisées un petit nombre de fois et un certain nombre de pages visitées un nombre moyen de fois.

Les pages du site, dont la fréquence est la plus faible, s'interprètent de deux façons : il s'agit soit de pages nouvellement introduites dans le site, soit de pages qui peuvent s'analyser en terme de «bruit» au sens statistique du terme. En effet, ces sommets risqueraient de perturber la lisibilité du réseau s'ils étaient conservés. Symétriquement, les pages, dont la fréquence d'apparition est forte, correspondent à une information «triviale» au sens où, étant trop génériques, elles ne permettent pas de discriminer l'information du corpus.

Lhen et al. (1995) ont montré que la courbe de Zipf peut se décomposer en trois parties à partir de la notion d'entropie de Renyi.

L'entropie d'ordre a, H_a, est définie par : $H_a = \frac{1}{1-a} \times \log \sum_{i=1}^{n} (p_i)^a$, telle que a soit différent de 1 et où n représente le nombre de modalités distinctes sur l'ensemble du corpus, et p_i la probabilité ou fréquence d'apparition de la modalité i dans le corpus.

Ces auteurs ont montré que l'entropie d'ordre 2 permet de déterminer le seuil séparant le trivial de l'intéressant et que l'entropie d'ordre ½ permet de séparer l'intéressant du bruit.

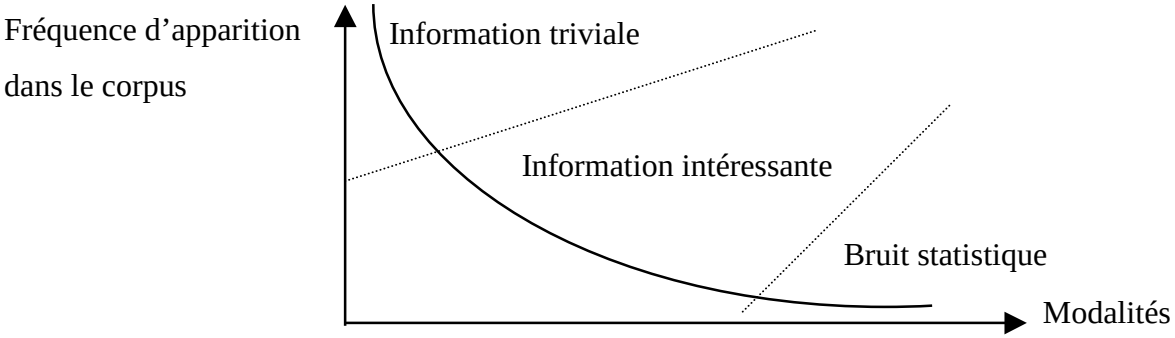

Figure 2 : Découpage de la courbe de Zipf en trois parties en utilisant l'entropie

A/ Le filtrage des paires.

Le point de départ de cette analyse consiste à remarquer que le réseau initial est inextricable en raison du trop grand nombre de liens qui y figurent. Or, ces liens n'ont pas tous le même poids. Les plus forts, qui correspondent aux arcs les plus épais, sont associés à des enchaînements de pages plus employés que d'autres. Si nous souhaitons nous intéresser aux liens entre les pages les plus souvent utilisés par les visiteurs, nous allons appliquer un filtre qui éliminera les liens les plus ténus. Les liens de fréquence inférieure à 6 correspondent au bruit statistique défini précédemment et ont été supprimés dans le réseau figure 3.

Dans chaque boîte figure le nom de la page suivi du nombre de visiteurs différents qui l'ont visualisée. Ce réseau fait apparaître deux activités sur le serveur : une activité ludique (entertainment/…) représentée sur la partie gauche du réseau et une activité recherche.

L'importance de la partie ludique, paradoxale sur un site consacré à la recherche, s'explique par l'existence de deux grands types de structuration de pages sur le serveur.

Le premier renvoie à une organisation linéaire. La page 1 possède un lien hypertexte avec la page 2 qui renvoie elle-même à la page 3. Lorsqu'il parcourt une telle page, l'utilisateur a deux alternatives : continuer ou revenir en arrière. L'autre type de structuration renvoie à une organisation en étoile. Dans ce cas, chaque page renvoie à de nombreuses pages potentielles si bien que l'utilisateur a un grand nombre de possibilités à chaque niveau.

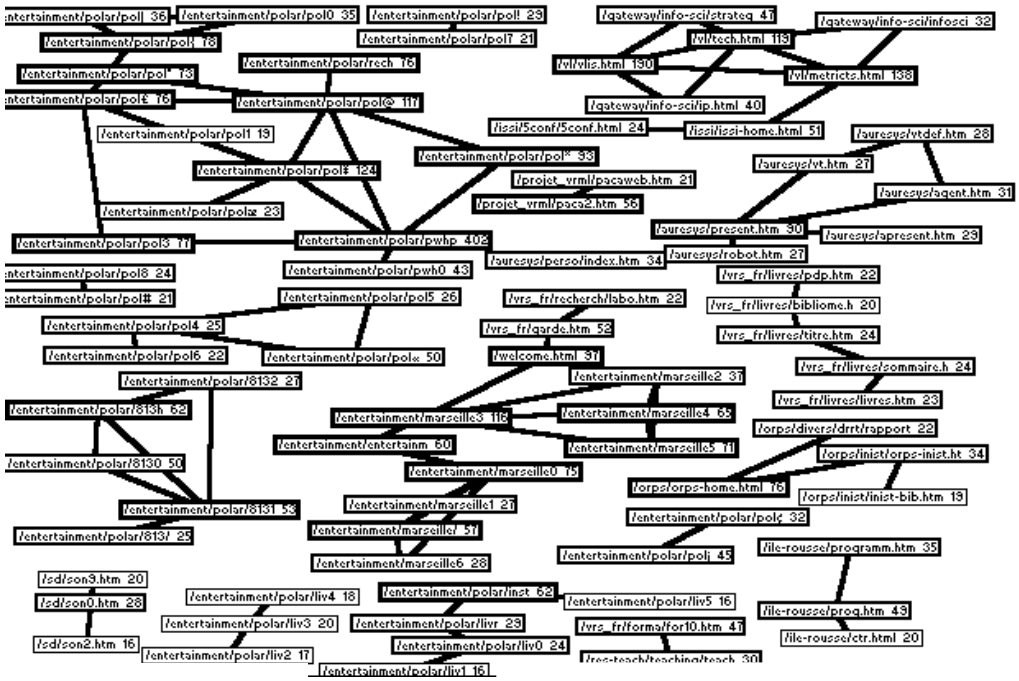

Figure 3 : Réseau obtenu en retenant les paires supérieures ou égales à 6.

Pour un même nombre d'individus connectés à la page 1, les individus qui se connecteront à la page 2 seront plus nombreux dans le premier cas que dans le second. La suppression des liens les plus faibles peut faire disparaître les groupes correspondant aux structures en étoile. Le fait de seuiller les paires survalorise les pages construites selon un mode linéaire. Dans le cas qui nous intéresse, la partie ludique est organisée de façon plutôt linéaire et la partie recherche de façon plutôt en étoile ce qui justifie cette sur-représentation.

Le choix d'un filtrage sur les paires n'est donc pas neutre sur la structure du réseau.

Il est possible d'extraire de ce réseau différents thèmes qui font l'objet de fréquentations spécifiques. Le visiteur se connecte souvent sur un espace très limité du site sans forcément élargir son investigation sur des activités connexes développées au CRRM. La relative étanchéité

entre les parties recherche et ludique peut se comprendre. Pourtant, on retrouve ce même phénomène au sein des composantes recherche de ce site. Il serait sans doute opportun de créer des passerelles adéquates entre les différents thèmes structurant la partie recherche.

B/ Le filtrage des formes.

Le point de départ de cette analyse est l'existence, au sein de l'ensemble des pages du site du CRRM, de pages plus visitées que d'autres. Pour accroître la lisibilité du graphe, nous allons retenir uniquement les pages les plus consultées. Ceci correspond à un filtrage des formes. Le réseau, figure 4, correspond au réseau initial pour lequel seules les pages visualisées 35 fois ou plus ont été retenues. Pour des raisons similaires, nous avons éliminé les paires de fréquence égale à 1. La caractéristique de ce filtrage est d'accorder plus d'importance aux sommets qu'aux liens.

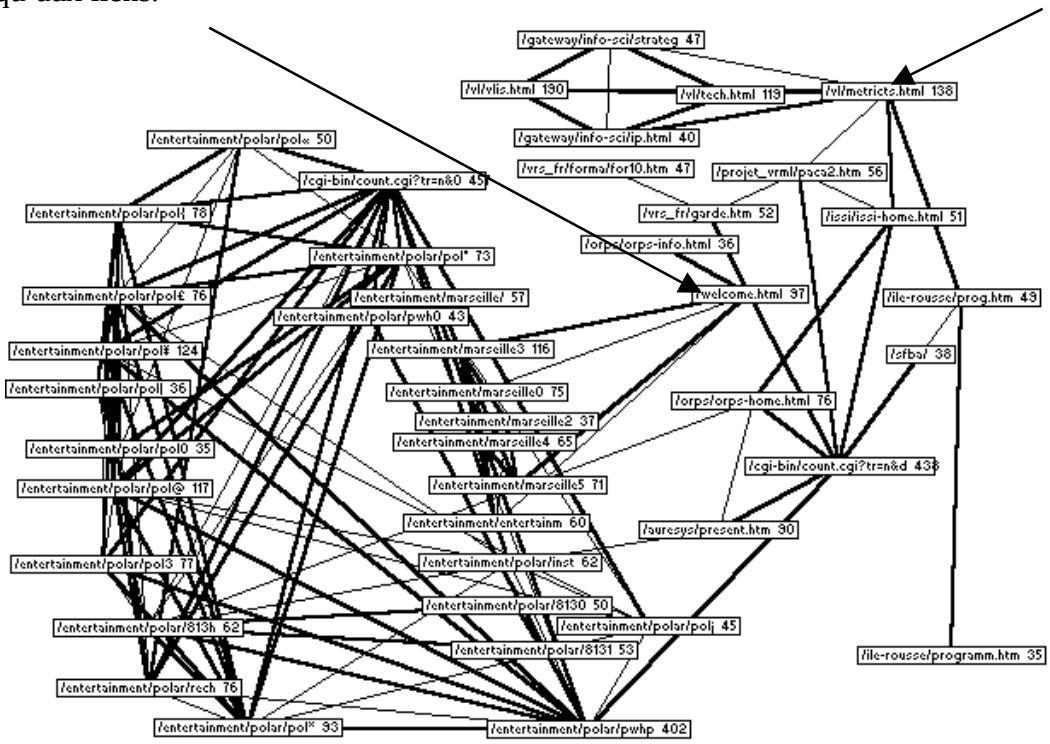

Figure 4 : Réseau[10] des pages visualisées plus de 35 fois lors des 2869 connexions.

Ce réseau confirme l'idée d'une segmentation entre partie recherche et partie ludique. Le lien entre ces deux composantes s'effectue par l'intermédiaire de la page "Welcome.html", qui n'a été consultée que 97 fois lors des 2869 connexions. Cette page est un point d'articulation du réseau (ou «cutpoint» dans la littérature anglo-saxonne). Celui-ci désigne un sommet du graphe dont la suppression permettrait d'augmenter le nombre de composantes fortement connexes du réseau. De même, le graphe fait ressortir le rôle tout à fait central de la page "vl

/metrics.html" qui apparaît au nord du réseau. Cette page est utilisée comme aiguilleur par les visiteurs du site dans la mesure où elle permet de renvoyer à d'autres pages. Elle est sans doute un modèle à suivre.

C/ Le filtrage des connectivités.

Un troisième filtre peut être appliqué pour rendre compte de la visite du site du CRRM. Il se fonde sur le fait que la grande complexité du réseau est souvent due à un petit nombre de sommets qui ont un grand nombre de relations avec beaucoup d'autres. On appelle connectivité associée à un sommet x, le nombre d'arcs partant de ce sommet.

Deux réseaux distincts peuvent être générés à partir de cette notion. Le premier consiste à éliminer les sommets desquels partent le plus de liens, et ainsi à mettre en avant la structure émergeante. Le second revient au contraire à ne retenir que les pages les plus liées aux autres, permettant ainsi de dégager la métastructure du réseau. Une telle approche revient à déterminer les sommets les plus centraux au sens de la centralité de degré[11].

C'est ce que nous avons fait dans la figure 5. Ce réseau retient les pages en relation avec au minimum 9 autres. Cet ensemble a la propriété d'être un 9-noyau. Il est composé de 14 pages. Si une seule de ces 14 pages est supprimée, alors au moins une des 13 pages restantes n'aura plus que 8 liens au total avec les autres.

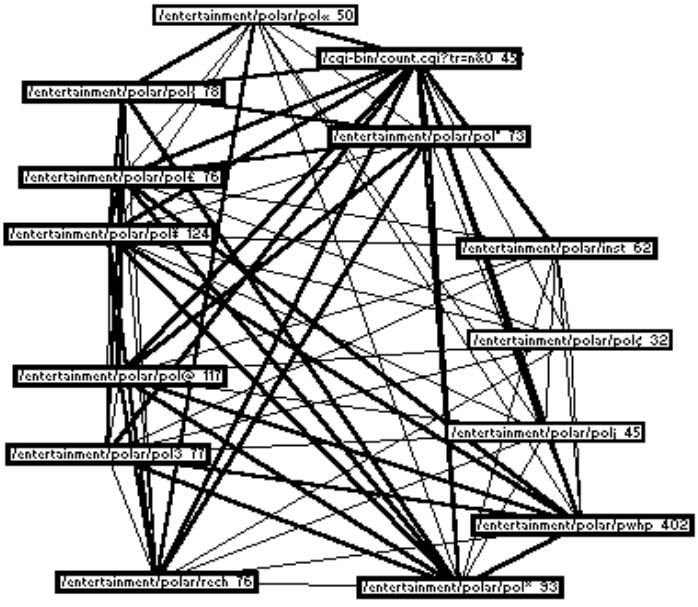

Figure 5 : 9-Noyau

Parmi les 2869 connexions analysées, il n'est pas possible d'identifier un k-noyau[12] avec k supérieur à 9. Ce 9-noyau a été obtenu par un processus itératif jusqu'à l'épuisement du filtre

sur les connectivités. Si on avait recherché un 10-noyau, aucun réseau n'aurait pu être représenté. Le 9-noyau définit donc un groupe de sommets qui entretiennent ensemble des relations denses. Ces sommets renvoient tous à des pages ludiques.

2. Vers une meilleure compréhension du comportement séquentiel des visiteurs du site.

Il est possible de construire des réseaux particuliers qui permettent d'apporter une réponse aux trois questions suivantes :

- **Quel est le degré de finesse de l'analyse** ? Chacune des analyses peut être conduite au niveau de l'ensemble des pages visitées, d'un groupe de pages correspondant à un thème particulier ou d'une page spécifique.
- **Quelle est la partie de la connexion qui nous intéresse** ? On peut prêter attention au point d'ancrage de la personne connectée, au point de rupture de la connexion, à une page particulière ou à l'ensemble des pages visualisées lors de la connexion.
- **Quelles sont les catégories de visiteurs qui nous intéressent** ? On peut étudier tous les visiteurs du site ou certaines catégories d'entre eux sur la base du temps passé sur le site ou de tout autre critère. Si le temps passé sur le site est un indicateur du degré d'intérêt manifesté par le visiteur, on pourra s'intéresser à tous les utilisateurs, à ceux qui se sont connectés pendant un long moment, aux visiteurs qui ont eu un parcours rapide sur le site ce qui peut dénoter une non-réponse à leur besoin.

A/ Le début et la fin de la connexion.

Deux moments privilégiés peuvent être objet d'analyse. Le premier porte sur la première page que visualise le visiteur lors de sa connexion. Le second s'intéresse à la page sur laquelle il va quitter le site. Plutôt que de les présenter successivement, ce qui reviendrait à étudier des tableaux de tri à plat[13], nous allons nous intéresser au premier et au dernier lien effectué par l'utilisateur. Dans les deux cas, pour des raisons de clarté, seules les paires de fréquences supérieures ou égales à trois ont été retenues, ce qui revient à survaloriser la partie ludique comme nous l'avons déjà mentionné.

Mille cinq cent neuf visiteurs ont cliqué au moins une fois sur un lien hypertexte sur le serveur du CRRM. Le réseau, figure 6, restitue donc environ 53 % de l'information initiale.

L'autre moitié des visiteurs du site du CRMM a eu un parcours qui se résume à l'arrivée sur une page et au départ immédiat sans avoir visualisé d'autres pages. Cette faible profondeur des visites est-elle spécifique au serveur du CRMM ou générale aux serveurs de recherche ?
Ce réseau permet de regrouper les différentes pages en huit composantes fortement connexes. La partie ludique, la plus représentée sur ce réseau, se subdivise elle-même en plusieurs composantes étanches. Les deux principales sont formées à partir des deux racines "entertainment \polar" et "entertainment\marseille".

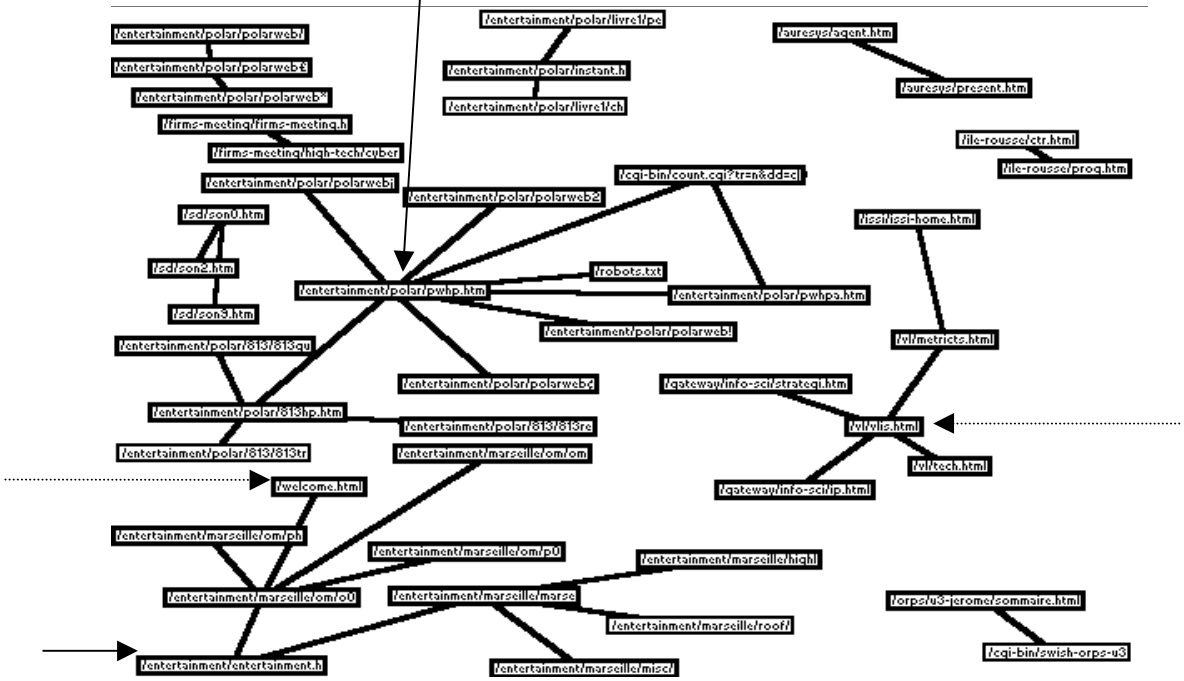

Figure 6 : Réseau construit à partir du premier lien de chaque visiteur.

La page d'accueil du CRMM, "welcome.html", utilisée seulement six fois en tant que telle par les visiteurs, soit 0.2% des connexions réalisées, apparaît tout à fait périphérique sur le réseau. Pourtant, le site est structuré autour de cette page qui a une double fonction : servir d'aiguillage et respecter une démarche pédagogique de présentation. La constatation du désintérêt pour la page d'accueil prédéfinie par les développeurs conduit à des réflexions relatives aux causes de ce phénomène et plus fondamentalement à ses implications.
Deux raisons peuvent être invoquées pour rendre compte du faible attrait de la page d'accueil. Tout d'abord, les visiteurs du site du CRMM bien souvent n'arrivent pas sur ce site pour la première fois. Ils ont alors pu recueillir ces pages dans un répertoire d'adresses et ainsi y accéder directement. La seconde raison est liée au fait que les pages, dont la fréquence d'apparition est relativement forte, correspondent souvent à des renvois directs à partir de moteurs de recherche ou d'autres sites. La page "vlis.htm" est par exemple un renvoi de la bibliothèque virtuelle du CERN.

Il existe donc à coté de la structure officielle du site, voulue et créée par l'équipe du CRMM, une structure émergente informelle qui tire sa légitimité de son utilisation. Son existence et son faible chevauchement avec la structure formelle conduisent à se poser quelques questions légitimes.

- Les développeurs du site du CRRM peuvent-ils continuer à structurer leur serveur autour d'une page pivot alors que celle-ci n'a aucune légitimité réelle ?
- Comment gérer le fait qu'il n'y a pas une, mais plusieurs pages d'accueil ?
- Comment faire en sorte qu'un utilisateur qui arrive sur une page qui n'est pas la page d'accueil officielle puisse se «brancher» sur d'autres parties du site ?
- Un utilisateur qui aurait une stratégie d'exploration du site non conforme à celle qui a été convenue par les développeurs du site perd-il quelque chose ?
- Les liens hypertextuels, si simples pour passer d'un concept à l'autre, n'échappent-ils pas aux développeurs du site dans le sens où l'information peut être lue dans n'importe quel sens?

La réponse à certaines de ces questions se trouve certainement dans la création de passerelles, liens horizontaux entre les différents thèmes abordés par le site.

Si on s'intéressait au dernier lien réalisé par le visiteur, on confirmerait et complèterait les conclusions précédentes : la partie recherche est minoritaire dans les pages d'abandon du site, les parties ludique et recherche sont disjointes et structurées autour de quelques groupes forts.

Il est cependant difficile d'aller beaucoup plus loin dans l'interprétation de ce type de résultats, parce que les raisons qui motivent la déconnexion du client peuvent être la satisfaction d'avoir obtenu l'information recherchée ou au contraire le dépit de n'avoir rien trouvé.

B/ Le degré de finesse de l'analyse.

Jusqu'à présent, nous avons retenu dans l'analyse toutes les pages du site du CRRM sans restriction de thème. Nous proposons maintenant d'étudier de façon spécifique le comportement des visiteurs qui sont passés lors de leur visite sur la page "vlis.html", qui appartient à la partie recherche du site.

Le réseau, figure 7, montre l'intérêt d'une telle analyse. Il représente l'ensemble des pages visualisées par les visiteurs qui sont entrés directement sur le site à partir de cette page et qui ont cliqué au moins deux fois sur un lien hypertexte (seuil déterminé par l'entropie d'ordre ½). Ces visiteurs correspondent d'une part à ceux que nous avions identifié comme les personnes envoyées par le serveur du CERN et d'autre part à ceux qui avaient mis cette page

dans leur répertoire. Le réseau visualise l'ensemble des liens et toutes les pages regardées lors des 65 connexions sans aucune restriction.

Ces visiteurs se démarquent des autres parce qu'aucune page ludique n'apparaît dans leur exploration. Leur comportement est peu dispersé et le réseau arrive à rendre compte clairement de leur navigation. Seules 27 pages distinctes ont été regardées par ces 65 personnes.

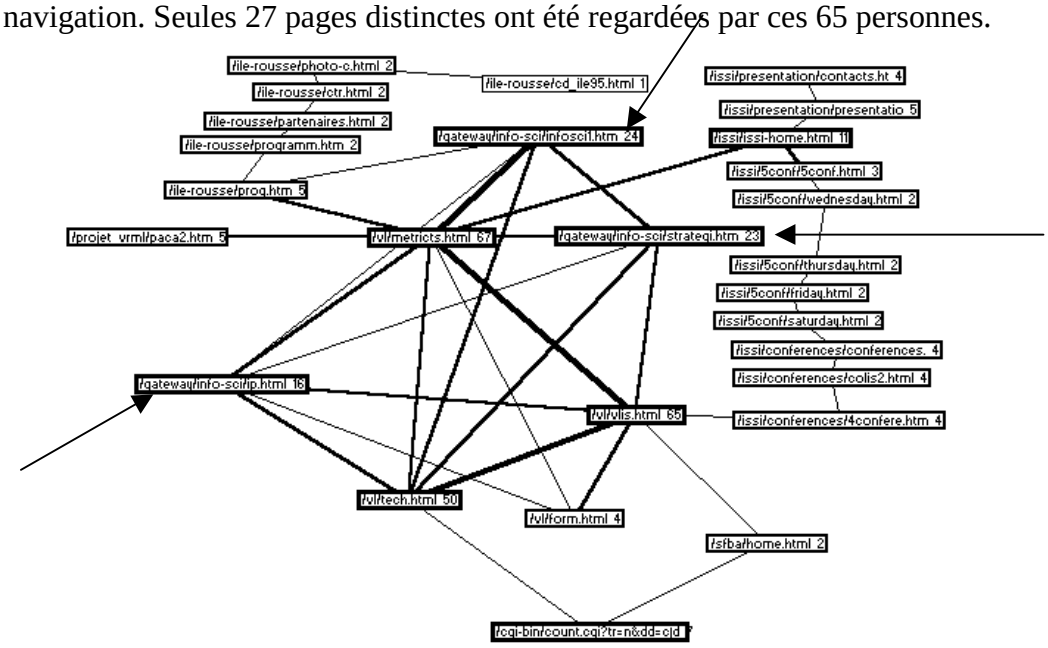

Figure 7 : Visualisation du parcours des 65 utilisateurs de la page «vl/vlis.htlm»

Malgré leur intérêt pour la veille technologique, les 65 visiteurs n'ont pas exploité toutes les informations du site du CRMM. Certains pans de la partie recherche comme tout ce qui concerne le robot "Auresys" ou "orps" sont absents de leur consultation. Sans doute une mise en évidence plus claire du lien vers la page d'accueil permettrait d'attirer ces visiteurs captifs vers les autres ressources du laboratoire.

En outre, le faible nombre de liens au total montre que la navigation de ces acteurs s'est cantonnée à quelques pages. Cette faible profondeur de lien tient au fait que les deux boîtes contenant le terme "gateway" correspondent à des ouvertures vers d'autres sites. Quarante-sept des 65 personnes connectées sont sorties du site par l'intermédiaire de ces passerelles qui apparaissent comme autant de fuites du système.

Le visiteur, qui consulte la branche vlis du serveur, entre dans une stratégie très particulière de recherche d'informations où le CRMM va jouer le rôle d'aiguilleur vers les autres ressources d'Internet.

C/L'analyse spécifique de certains visiteurs.

Nous allons étudier les visiteurs du site du CRMM en fonction de la durée de leur connexion. Nous avons choisi de privilégier les 169 visiteurs qui se sont connectés plus de dix minutes. Ces visiteurs, déterminés par le calcul de l'entropie d'ordre ½, correspondent au bruit défini précédemment. Ce type d'analyse spécifique se justifie par le fait que l'on peut penser que les personnes qui se sont connectées le plus longtemps sur un site sont celles qui lui manifestent le plus d'intérêt. Elles méritent que l'on s'intéresse à elles de façon prioritaire.

Leur analyse a été appréhendée par deux réseaux successifs. Le premier, figure 8, visualise les fréquences de paires supérieures à quatre. Le second, figure 9, s'intéresse aux pages consacrées à la recherche visualisées neuf fois ou plus lors de ces 169 visites.

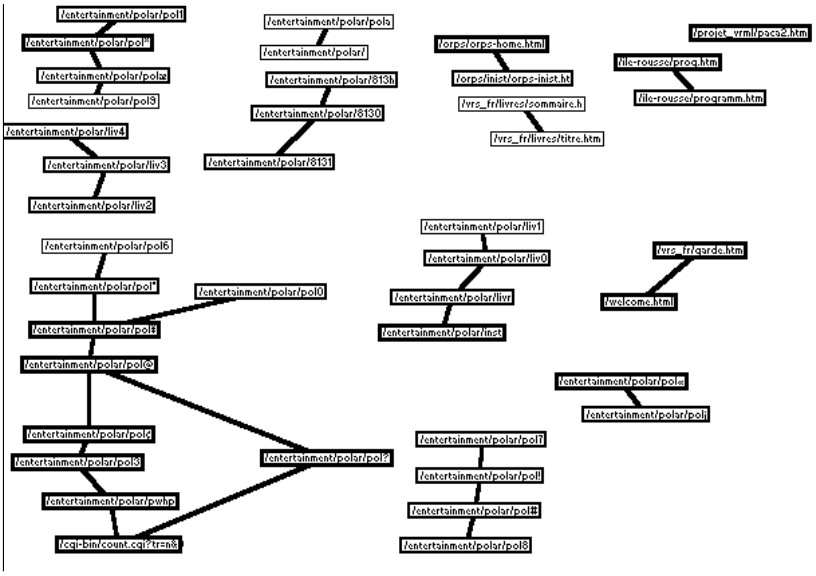

Figure 8 : Réseau des utilisateurs restés plus de dix minutes sur le site du CRRM en retenant les fréquences de paires supérieures à 4.

La figure 8 confirme les intuitions précédentes. La présence des pages ludiques sur ce réseau est forte chez les visiteurs qui se connectent plus de dix minutes. Dans cette composante ludique, la partie "entertainment\polar" occupe une place essentielle. Cette position privilégiée correspond à un site consacré au roman policier fortement plébiscité par les visiteurs, donc certainement extrêmement pertinent.

L'activité ludique de ce serveur n'est pas seulement une activité complémentaire de la partie recherche. Cette partie consacrée aux romans policiers est à elle seule un point d'attraction du site. 243 ont comme point d'entrée sur le site la page "pwhp.htm" traitant ce thème.

La part consacrée à la recherche est mineure sur ce réseau, mais le zoom, figure 9, permet de commenter les axes de recherche du site les plus consultés. On note une forte interrelation entre ces différents axes structurés de façon quasi arborescente autour de la page d'accueil.

Ce réseau montre que les visiteurs, qui se sont connectés pendant plus de dix minutes et qui s'intéressent à la recherche, ont une démarche de recherche systématique orienté vers l'exploration en profondeur du site du CRMM.

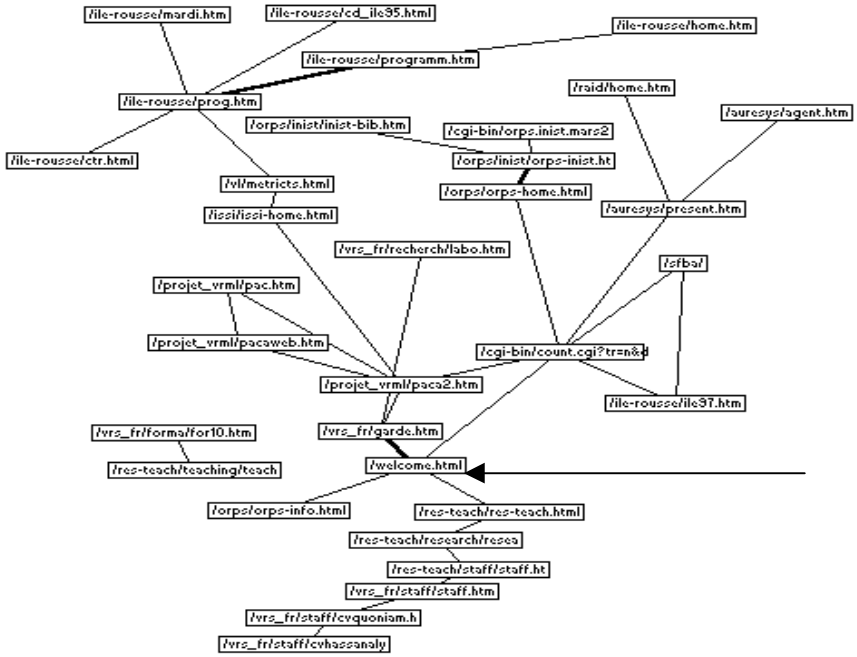

Figure 9 : Réseau représentant les pages consacrées à la recherche visualisées neuf fois ou plus par les visiteurs connectés plus de dix minutes.

III LES APPORTS DE L'ANALYSE RESEAU.

L'analyse de réseau a permis de dépasser l'information fournie par les analyseurs commerciaux de fichiers .*Log* et devrait offrir aux entreprises de nouveaux moyens pour mesurer l'audience de leur site Internet.

1. Les apports de l'analyse réseau par rapport aux analyseurs commerciaux de Log.

A travers cette recherche, nous avons pu déterminer les points de passage obligés du visiteur lorsqu'il se connecte au site du CRRM, le parcours de l'utilisateur type, l'articulation entre les différentes thématiques du site, les pages obsolètes, les pages d'accueil et de sortie.
Par rapport aux analyseurs du commerce, l'analyse réseau reconstitue la démarche par laquelle l'utilisateur a abordé et a quitté le site. Elle prend en compte la dimension séquentielle de la consultation. A même de créer des liens entre les informations disponibles, elle donne un sens

au parcours du visiteur. Elle rend possible la visualisation d'une page dans son contexte, sa caractérisation par un certain niveau de centralité.
De plus, l'approche réseau dynamise l'information disponible. Elle détermine les thèmes faisant l'objet d'une fréquentation particulière. Elle spécifie les pages fortement connectées les unes aux autres et ainsi la métastructure du serveur.
Enfin, nous avons pu définir un certain nombre de réseaux selon le degré de finesse de l'analyse voulue, la partie de la connexion qui nous intéressait et caractériser les clients du serveur sur la base du temps passé sur le site.

2. Les apports marketing de l'analyse réseau.

Sur Internet, le marketing se fonde sur l'initiative primordiale laissée au client à la recherche d'informations voire d'achats et sur la réaction de l'entreprise à ses demandes et requêtes. Le client communique, l'entreprise écoute et satisfait ses besoins. Son problème est de disposer en permanence de moyens pour créer un «plus» pour ses visiteurs en introduisant des informations utiles et fiables à chaque étape du processus de décision et d'achat.
Pour y arriver, elle a besoin d'un site attractif en termes de contenu de l'information, de convivialité, de facilité de navigation et même d'interactivité. L'analyse réseau des fichiers .*Log* est à même d'apporter des réponses sur la pertinence des choix de l'entreprise pour les trois premiers critères.
Cette analyse a permis de mieux appréhender le comportement des visiteurs en matière de navigation sur le site. Il pourrait en être de même sur leur manière de passer des commandes, si nous avions analysé un site marchand. Un tel outil, grâce à la connaissance qu'il offre, doit aider l'entreprise à accroître le panier moyen d'achat sur son site en montrant les forces et les faiblesses de l'interface.
En outre, il est possible de différencier les visiteurs en fonction de leur utilisation du site, de segmenter par exemple les acheteurs et les non-acheteurs sur un site marchand. On peut déterminer le parcours de chacun sur le site, les pages qui les intéressent, les services qui les attirent.
Comme cette approche permet de construire les réseaux à partir d'une page particulière, une mesure de l'influence des offres promotionnelles de têtes de gondoles virtuelles est réalisable.

En résumé, grâce aux filtrages réalisés, l'analyse réseau indique le niveau de pertinence d'une page d'un site Internet. Quels que soient le niveau d'analyse, son degré de finesse, la partie de la connexion ou la catégorie de visiteurs qui nous intéresse, ces filtrages sont nécessaires.
Une page sera pertinente si

- lors d'un filtre sur les paires, il existe des liens forts entre cette page et les autres,
- lors d'un filtrage des formes, elle est visualisée un grand nombre de fois,
- lors d'un filtre sur les connectivités, elle possède un grand nombre de liens vers d'autres pages.

Par exemple, lorsque l'entreprise s'interroge sur la page sur laquelle elle devrait mettre une publicité, le filtrage des connectivités semble particulièrement intéressant. En effet, l'obtention du k-noyau du réseau permet de connaître les pages qui entretiennent des relations denses entre elles et le thème qu'elles abordent. Une publicité sur ce thème aura plus d'impact si elle est placée sur une de ces pages en raison de l'audience qu'elle sera assurée d'avoir : tout visiteur visualisant une page quelconque du k-noyau passera systématiquement sur la page où se trouve la publicité. Ceci lui permettra d'affiner son ciblage et de réduire son coût de contact utile : par rapport à la page la plus visualisée sur ce thème, page pour laquelle on n'a pas d'idée précise sur les raisons de sa fréquentation, l'entreprise est sûre ici de toucher les personnes particulièrement sensibles à ce thème.

Deux voies de recherche futures se présentent.
1. Nous avons dans le cadre de cette recherche amélioré la compréhension du comportement séquentiel des utilisateurs sur un site Internet. Toutefois, il ne nous est pas possible d'appréhender les fondements de leur comportement. Il semble donc important de mettre en place des plans expérimentaux en laboratoire pour améliorer la connaissance du mode de fonctionnement de l'utilisateur devant un site commercial et pour saisir les raisons de son comportement.
2. La distribution zipfienne, qui nous a permis de décomposer l'information disponible en trois parties, correspond à un système dynamique équilibré. A un instant t, un corpus est composé d'un continuum allant de modalités émergentes à faible fréquence à des modalités «mûres» pour lesquelles la fréquence d'apparition est plus forte. A l'instant t+1, se produit une translation de la courbe vers le haut tandis que de nouvelles modalités émergentes apparaissent. Grâce à ce système dynamique, nous devrions pouvoir évaluer les modifications de la

visite d'un site au cours du temps et ainsi mesurer l'efficacité des actions promotionnelles ou des transformations réalisées.

Dans ce type d'étude, nous avons volontairement choisi d'appliquer la méthode réseau à un domaine d'étude extrêmement concurrentiel. Compte tenu de l'importance économique que revêt la mesure d'audience sur le Web, ce type d'application devrait se développer à un rythme rapide et un grand nombre de sociétés commerciales se positionnent déjà sur ce segment. L'étude de la concurrence ne nous a pas permis, à ce jour, d'identifier des outils qui soient positionnés sur le créneau de l'analyse réseau. Il s'agit donc pour l'instant d'une niche isolée à fort potentiel de développement.

Notes :

1. Le lecteur souhaitant approfondir ces notions pourra se référer à l'ouvrage de Stout (1997).

2. - Un exemple de ligne en Common Log Format est :
194.51.254.3 - - [01/Dec/1996:01:37:55 -0100] "GET /entertainment/polar/polarweb/pwtete.htm HTTP/1.0" 200 2440 où:
   - 194.51.254.3 désigne le nom de la machine du visiteur du site
   - [01/Dec/1996:01:37:55 -0100] l'heure de la connexion
   - "GET /entertainment/polar/polarweb/pwtete.htm HTTP/1.0" le nom de la page visualisée par le visiteur
   - 200: cette valeur indique si le fichier a été correctement trouvé (code 2xx)
   - 2440 désigne le nombre d'octets transférés.

3. Un proxy désigne un ordinateur qui s'intercale entre un réseau privé et l'Internet pour faire office de cache et qui enregistre les pages Web transférées par les utilisateurs pour les délivrer sans qu'il soit nécessaire de se connecter sur le serveur initial.

4. Les analyseurs de Log, que nous avons testés, ont été téléchargés en version d'évaluation à partir d'Internet

5.Analog est un logiciel créé par Stephen R.E. TURNER du «Statistical Laboratory, University of Cambridge», sret1@cam.ac.uk. Logiciel téléchargeable à l'adresse suivante : http://www.statslab.cam.uk/~sret1/analog/

6.Websracker 21.42 est un logiciel créé par la société «Cambridge Quality Management» Inc. 1639, 9th Avenue, San Francisco, California 94122 USA. Une version d'évaluation de ce logiciel est disponible à l'adresse suivante : http://www.CQMInc.com/

7.Hitlist est un logiciel créé par la société «Marketwave Corporation» 1415 Western Avenue, suite 488, Seattle WA USA 98101. Une version d'évaluation peut être téléchargée à l'adresse: http:///www.marketwave.com/

8. Une session désigne l'ensemble des pages parcourues par un visiteur lors d'une visite sur le site

9. Une visite peut se définir comme l'ensemble des requêtes effectuées par un visiteur lors d'une connexion

10. Les «/cgi-bin/count.cgi?tr=n&d» ne sont pas des pages mais des compteurs qui s'incrémentent lorsque la page où ils sont situés est lancée. Pour cette raison, nous avons négligé ces cellules du réseau.

11.Au sens de la centralité de degré, un sommet i est plus central qu'un sommet j s'il a plus de sommets qui lui sont adjacents que le sommet j. Mathématiquement, la centralité de degré associée au sommet i s'obtient par la formule suivante : $cd'_i = \sum_{j=1}^{n} xij$ où x_{ij} désigne une valeur binaire égale à 1 s'il existe un arc entre les sommets i et j et 0 dans le cas contraire.

12. Un k-noyau regroupe un ensemble de sommets tels que chaque sommet est relié directement à au moins un nombre k d'autres sommets du même groupe.

13. Ces tableaux ont été réalisés et nous ferons parfois référence dans le texte à leur contenu. Nous ne les avons pas exposés dans notre développement, parce que nous souhaitions uniquement nous concentrer sur les apports de l'analyse réseau.

Annexe 1 : : Les axes de recherche présents sur le site du CRRM.

/vl/metricts.html : Librairie Virtuelle en sciences de l'information animée par Luc Quoniam

/issi/issi-home.html : The International Society for Scientometrics and Informetrics organise des biennales. La prochaine a lieu à Jérusalem en juin 1997. Le CRRM est présent à chacun de ces colloques.

/orps/orps-home.html : L'Observatoire Régional de la Production Scientifique Provence-Alpes-Côte-d'Azur est réalisé par le CRRM avec la collaboration d'un certain nombre d'Institutions et de personnes. Il doit permettre une meilleure lisibilité de l'ensemble des productions scientifiques des différents acteurs régionaux.

/ile-rousse/prog.htm : Ce Colloque est organisé par la Société française de bibliométrie appliquée (SFBA). Il se tient tous les deux ans en Corse autour des thématiques suivantes : Bibliométrie, Linguistique, Information Stratégique, Veille Technologique, Intelligence Économique.

Ce site comprend la présentation interactive du colloque et de la SFBA, les textes intégraux des communications, les photos du colloque et des participants.

/auresys/present.htm : La conception d'un robot de recherche automatique est l'objet d'un mémoire de DEA Information Scientifique et Technique effectué au CRRM par Bruno Mannina. AURESYS permet de créer des Bases de Données interrogeables à distance par plusieurs utilisateurs. Les informations sont récupérées du NET, puis traitées pour être stockées dans ces bases. Chaque Base de Données répond à une requête personnalisée d'un utilisateur. Grâce aux Bases de Données créées, AURESYS donne les moyens à un utilisateur d'avoir un maximum de renseignements se rapportant à un sujet donné. Constitution d'un corpus analysable directement par des outils bibliomètriques.

/projet_vrml/paca2.htm : Ce projet du CRRM fait l'objet de deux axes d'investigation :

- Le premier développé par Pascal Faucompré permet de faire un lien Sciences, Technologies.

En effet la base, issue de la base PASCAL (©INIST), contient essentiellement des références scientifiques et il nous est apparu intéressant de lier ces références avec un aspect technologie et propriété industrielle.

- Le second permet d'interroger la base sur des critères purement statistiques. Il a été possible de fabriquer une représentation utilisant la réalité virtuelle permettant d'appréhender l'analyse avec une interface graphique.

Bibliographie.

AHO A., HOPCRAFT J. et ULLMAN J. (1987), *Structure des données et algorithmes*, Interéditions.

BOURGNE P. (1997), Etat de l'utilisation de la notion de réseau en marketing, *Actes de l'Association Française de Marketing*, Bordeaux, 14, 437-456.

BOUTIN E. (1997), *Analyse du Log de la technopôle de l'Arbois*, Papier de Recherches, Laboratoire Le Pont, Mars.

BOUTIN E. (1999), *Le traitement d'une information massive par l'analyse réseau : méthode, outils et applications*, Thèse de Doctorat en Sciences de l'information, soutenance prévue janvier.

BOUTIN E. et FERRANDI J.M. (1996), La construction automatique de réseaux sociaux : Etude exploratoire, *Actes du Colloque National de Recherche en IUT en Mathématiques, Statistiques, Informatique et leurs applications*, Clermont-Ferrand.

BOUTIN E., FERRANDI J.M. et VALETTE-FLORENCE P. (1996), Les réseaux comme outil d'analyse des chaînages cognitifs : une illustration expérimentale, Papier de recherche, 96-06, publié au Centre d'Etudes et de Recherches Appliquées à la Gestion (CERAG), Université Pierre Mendes France, Ecole Supérieure des Affaires.

BOUTIN E., FERRANDI J.M. et VALETTE-FLORENCE P. (1997), L'analyse des chaînages cognitifs et la construction automatique de réseau comme outil de veille commerciale, *International Journal of Information Science for Decision Making*, 0, 19-34.

BOUTIN E., QUONIAM L., ROSTAING H. et DOU H. (1995), A new approach to display real co-authorship and co-topicship through network mapping, *Actes du « Fifth International Conference on Scientometrics and Infometrics »*, Chicago, 7-10 Juin 1995.

COSTES Y. (1998), La mesure d'audience sur Internet, *Décisions Marketing*, 14, 63-71.

DEGENNE A. et FORSE M. (1994), *Les réseaux sociaux: une analyse structurale en sociologie*, Editions Armand Colin.

IACOBUCCI D. (1996), *Networks in Marketing*, Sage Publications.

LHEN J., LAFOUGE T., ELSKENS Y., QUONIAM L. et DOU H. (1995), La statistique des lois de Zipf, *Actes du Colloque, Les systèmes d'informations élaborés*, Ile Rousse.

ONNEIN-BONNEFOY C. (1997), Les bandeaux publicitaires sur Internet : mesures d'efficacité, *Décisions Marketing*, 11, 87-92.

REBOUL P. et XARDEL D. (1997), *Le commerce électronique*, Editions Eyrolles.

ROSTAING H. (1993), *Veille technologique et bibliométrie : concepts outils et applications*, Thèse de Doctorat en Sciences de l'Information, Université des Sciences et des Techniques d'Aix- Marseille III.

SCOTT J. (1991), *Social Network Analysis*, Newbury Park CA, Sage.

STOUT (1997), *Web Site Stats : Tracking Hits and Analysing Traffic*, Osborne McGraw-Hill.

TAGUE ET NICHOLLS (1987), The Maximal Value of a Zipf Size Variable : Sampling Properties and Relationship to Other Parameters, *Information Processing and Management*, 23, 3, 155-170.

WASSERMAN S. et FAUST K. (1994), *Social Network Analysis*, Cambridge, Cambridge University Press.

ZIPF G.K. (1949), *Human Behavior and the Principles of Least Effort*, Addison Wesley.