Audit d'un serveur Internet et approche réseau

Eric Boutin, Hervé Rostaing, Luc Quoniam

AUDIT D□UN SERVEUR INTERNET ET APPROCHE RESEAU

Eric Boutin *
Hervé Rostaing **
Luc Quoniam **

* Laboratoire Lepont, Université de Toulon et du Var IUT TC BP 132 83957 La Garde cedex, boutin@univ-tln.fr
** Laboratoire CRRM, Centre de recherche Saint Jérome 13397 Marseille cedex 20

AUDIT D'UN SERVEUR INTERNET ET APPROCHE RESEAU.

Grâce au WEB, une entreprise, une Université, une organisation a la possibilité de mettre à la disposition de la communauté sur internet une information qui se présente sous forme de pages au format html qui s'enchaînent par un processus hypertexte. L'objectif recherché est, suivant le cas, de constituer une vitrine de ses activités, de répondre à une stratégie d'image ou de prestige, d'appuyer une stratégie commerciale ou d'être présent sur un nouveau médium.
Un serveur ainsi constitué est consulté par des utilisateurs aussi appelés clients.
Il est possible d'appliquer à cette confrontation entre l'offre et la demande d'information une des règles élémentaires du marketing: l'offre doit s'adapter à la demande car un site n'est pertinent que s'il est consulté.
Il est donc important pour l'organisation de disposer de capteurs dans son environnement qui lui permettent d'obtenir des informations statistiques correspondant à la visite de son site. Nous nous intéresserons ici à une source d'information incomplète mais toujours disponible. Elle résulte de l'utilisation du fichier Log qui enregistre les connexions des différents utilisateurs. L'analyse du fichier Log, sur une période donnée, permet de dégager des invariants et fournit les réponses aux questions suivantes:
- Quelles sont les pages les plus visitées?
- Sur quelles pages les clients restent-ils le plus?
- Quel est le temps moyen passé par un utilisateur lors d'une connexion?
- A partir de quelle page les utilisateurs arrivent-ils sur notre site?
- A partir de quelle page les utilisateurs quittent-ils notre site?
- Quels sont les thèmes de notre site qui sont les plus consultés?

Ces éléments statistiques fournissent de précieuses indications sur la façon dont le site est utilisé et permettent au site de s'adapter aux besoins.

Il existe un certain nombre de logiciels commerciaux, connus sous le nom d'analyseurs de Log qui, partant du fichier .Log, proposent un ensemble d'indicateurs statistiques d'utilisation du serveur.
Cette étude a pour objectif de renouveler l'approche traditionnelle des analyseurs de Log en utilisant l'analyse en terme de réseau (Degrenne et alii, 1994). Cette nouvelle démarche est présentée en s'appuyant sur l'exemple de l'audit du serveur du Crrm réalisé en Décembre 1996 à partir de 2869 connexions. Le Crrm est un centre de recherche, qui a construit un site Web permettant de présenter ses axes de recherche : produits, chercheurs, colloques, publications... D'autre part, le Crrm permet à des étudiants de troisième cycle qui souhaitent s'exprimer, d'héberger sur son serveur leur pages html. Les thèmes de ces pages ne touchent pas forcément au domaine de la recherche.

Dans un premier temps, nous positionnerons la méthode réseau par rapport aux méthodes existantes.
Dans un second temps, nous développerons un bref aperçu de la richesse des analyses qui peuvent être conduites par cette méthode.

I- POSITIONNEMENT DE LA DEMARCHE RETENUE:

Nous présenterons dans cette partie les étapes séquentielles conduisant de l'information brute, contenue dans le fichier Log, à une information statistique élaborée. L'analyse successive de ces différentes étapes permet de dégager les spécificités de la méthode d'analyse en terme de réseau par rapport aux méthodes utilisées dans les analyseurs de .Log traditionnels. Dans un souci didactique, nous raisonnerons dans cette partie sur un exemple simple, constitué d'un extrait du fichier Log

correspondant à 3 des 2869 connexions réalisées sur le serveur du CRRM. Cet exemple sera élargi par la suite.

Différentes étapes sont nécessaires pour passer de l□information brute à une information de synthèse. Ces différentes étapes constituent les différents maillons d□une chaîne de traitement de l□information présentée figure 1. Nous allons les identifier successivement.

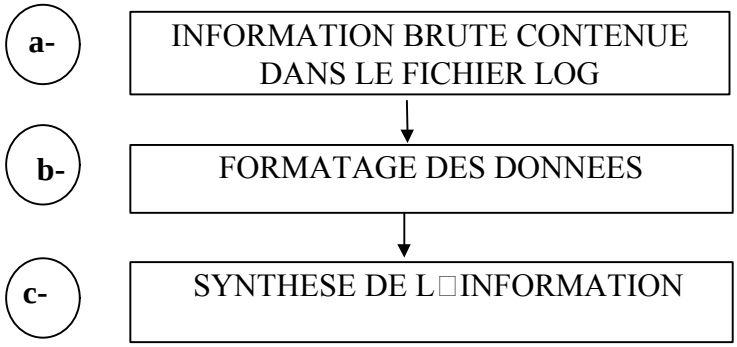

Figure 1:La chaîne de traitement de l'information

a- L'information brute: contenu et limites.

Les analyseurs de .Log reposent sur l□exploitation d□un fichier appelé Log fourni par le serveur. Ces fichiers. .Log sont structurés de la même manière. Nous allons nous attacher à présenter un exemple de fichier. .Log dans un double souci:

Présenter la nature des informations contenues dans ce fichier
Présenter les limites de cette information.

L□exemple que nous allons considérer correspond à 3 des 2869 connexions réalisées sur le serveur du CRRM du 01/12/1996 au 20/12/1996. Ce fichier contient l□information relative à la succession des pages visualisées par ces trois utilisateurs lors de leur connexion. Il se présente comme une succession de lignes. Ces trois utilisateurs ont visualisé respectivement 5, 3, 2 pages du serveur du Crrm comme le montre la figure 2.

```
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
crrm.univ-mrs.fr - - [01/Dec/1996:12:12:20 -0100] "GET /images/fond_gr2.gif HTTP/1.0" 200 9841
```

Une Connexion

Une page visualisée

Figure 2: Exemple de fichier Log.

α- Le contenu de l□information:
Ce type de fichier .log fournit trois informations principales:
- En début de chaque ligne figure le nom du serveur de la personne connectée. Ce nom joue le rôle d□identifiant de l□utilisateur. Le nom du serveur du premier utilisateur est par exemple « 194.51.254.3 ».
- Le fichier .Log contient le nom de la page visitée par le client. Ainsi, la troisième page visitée par le premier client est « / entertainment/polar/polarweb/album.htm HTTP »
- Enfin, chaque connexion à une page donnée est datée.

β- Les limites de cette information:
Cette information doit être interprétée avec précaution pour des raisons tenant au mode d'identification de l'utilisateur et au mode de construction du fichier .Log.
- Limite liée au mode d'identification de l'utilisateur. Chaque utilisateur n'est pas identifié de façon univoque dans le fichier. Si deux utilisateurs ayant le même nom de serveur se connectent au même moment, il ne sera pas possible de désambiguiser ces deux utilisateurs. Ceci introduit un risque de confusion potentielle. D'autre part, l'adresse du serveur ne permettra jamais de connaître le nom de l'utilisateur ni son adresse électronique.
- Limite associée au mode de construction du fichier .Log. Le fichier .Log enregistre toute **nouvelle** page visualisée par le client du site. Il est donc impossible de retranscrire l'intégralité du cheminement d'un utilisateur sur un site donné: l'enchaînement de deux pages dans un fichier .Log ne signifie pas toujours que les deux pages en question sont liées l'une à l'autre par un lien hypertexte direct.

b- Le formatage des données:

L'étape du formatage va consister à structurer l'information selon un format qui facilite l'exploitation future des données. Il est nécessaire d'automatiser la procédure de reformatage en utilisant une routine informatique pour deux raisons principales. La première est qu'un fichier .Log comprenant plusieurs milliers de lignes, il est impossible de concevoir un traitement manuel d'une telle masse de données. En second lieu, les fichiers .Log présentent une structure homogène si bien qu'une procédure automatisée pourra être réutilisée. Une routine a été développée à cet effet. Elle permet, à partir des trois indicateurs de base fourni lors d'une connexion (nom du serveur du client, date de la connexion, nom de la page visualisée), de créer par combinaison 5 indicateurs supplémentaires. Le type de reformatage présenté ci dessous ne prétend pas à l'universalité ni à l'exhaustivité. Toutefois, à partir de trois données de base, il n'est pas possible de multiplier à l'infini les indicateurs potentiels. La structure de l'information reformatée à partir des 5 premières lignes du fichier précédent est présentée dans le tableau 1.

```
-- -
Adresse:194.51.254.3
profondeur: 1
heure:01/12 01:36:46
relations:ND
somtps:-9
tps:0
last:non
-- -
Adresse:194.51.254.3
profondeur: 2
heure:01/12 01:37:55
relations:/cgi-bin/Count.cgi?tr=N&dd=C|df=polar.dat @/entertainment/polar/polarweb/pwtete.htm
somtps:-9
tps:1
last:non
-- -
Adresse:194.51.254.3
profondeur: 3
heure:01/12 01:38:28
relations:/entertainment/polar/polarweb/pwtete.htm @/entertainment/polar/polarweb/album.htm
somtps:-9
tps:1
last:non
-- -
Adresse:194.51.254.3
profondeur: 4
heure:01/12 01:42:09
relations:/entertainment/polar/polarweb/album.htm @/entertainment/polar/polarweb/pwcritik.htm
somtps:-9
tps:4
last:non
```

Tableau 1: Un exemple de fichier . .Log reformaté

Cette information est structurée autour de 4 références. Pour chacune d'elles sont définis 7 champs renseignés par une ou plusieurs formes. Le tableau 2 présente une synthèse du vocabulaire utilisé:

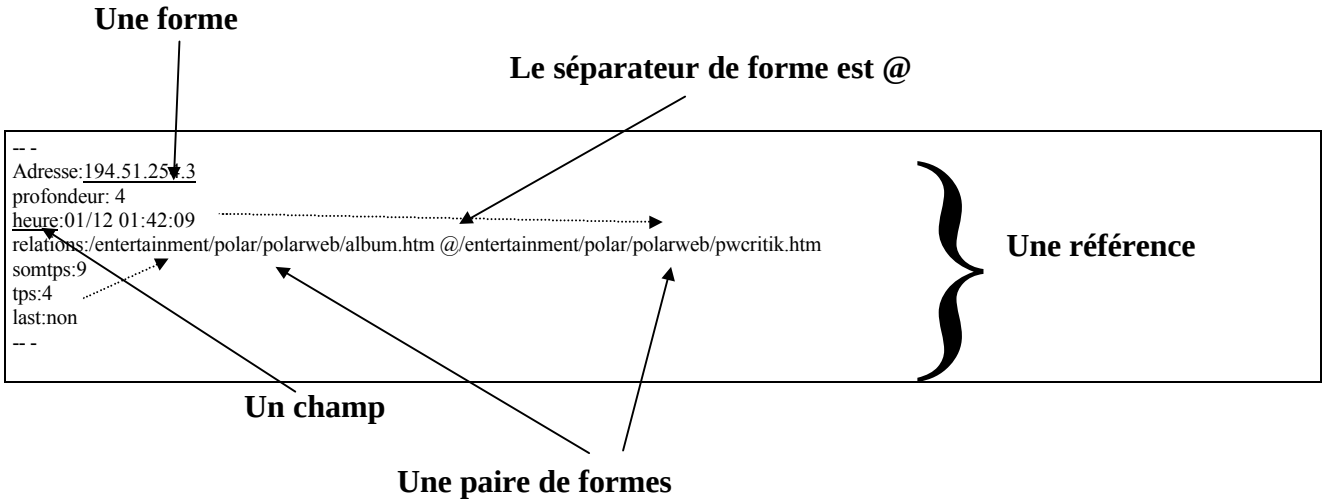

Tableau 2: Synthèse du vocabulaire utilisé.

Cette information est plus riche que l'information brute disponible initialement. En effet, en plus du nom du site, du nom de la page et de la date de la connexion apparaissent d'autres informations.
- Le niveau de profondeur correspond au nombre de pages successives que l'utilisateur a visualisé 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 l'utilisateur a « navigué » sur le site.
- Lorsque l'utilisateur passe d'une page à l'autre, cette information est recueillie dans la rubrique « relations » figurant dans la référence. Cette rubrique permet de passer d'une information statique contenue dans le fichier initial à une information dynamique. Pour un niveau de profondeur de 1, le champ relation est renseigné par la mention « ND » (Non Défini) car l'utilisateur vient de se connecter.
- En faisant la différence entre les dates prises deux à deux, pour une connexion donnée, il est possible de connaître le temps passé par un client entre deux pages. Cette information est contenue dans la rubrique « tps: » et est exprimée en minutes. Plus la valeur de cet indicateur est élevée:

* plus le temps passé sur la page elle même est élevé

et/ou

* plus le temps de transfert entre cette page et la suivante est élevé. Le temps de transfert entre deux pages 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 et du nombre de page intermédiaires à visualiser avant d'arriver sur une page nouvelle. Cet indicateur peut donc difficilement être interprété comme un indicateur de pertinence d'une page.
- En agrégeant le temps passé sur chaque page, on obtient le temps total passé par le client sur le site. Cette information est contenue dans le champ somtps lui aussi exprimé en minutes.
- On identifie lorsque le champ « last:=oui » la dernière page visualisée par un client avant qu'il ne quitte le site.

Il est possible de définir d'autres indicateurs en fonction de ce que l'analyse veut montrer: jour auquel a lieu la connexion, pays d'origine du client. Toutefois, nous n'avons pas retenu ces paramètres.

Ce processus de reformatage des données s'accompagne de l'élimination des données non pertinentes pour l'analyse. Elles sont de deux types:

- Elles correspondent tout d□abord aux connexions établies sur le site en interne. Dans la mesure où l□objectif est de faire l□audit d□un serveur, il peut être judicieux d□analyser de façon disjointe les connexions réalisées en interne des connexions réalisées à partir d□autres serveurs. C□est la raison pour laquelle ne figurent pas dans le fichier ci dessus les connexions correspondant au serveur « crrm.fr. »
- Lorsqu□une page se charge, chaque image associée à cette page fait l□objet d□une ligne supplémentaire dans le fichier .Log. Lorsqu□une page est riche en image, ceci surcharge le fichier .Log. Dans la mesure où l□objectif est de suivre les actions effectuées par l□utilisateur, il ne faut pas prendre en compte ces divers chargements qui sont automatiques. C□est la raison pour laquelle, dans notre exemple, les pages ou figurait une extension « .gif » au fichier ont été supprimées.

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

c- La synthèse de l'information:

Il est possible de valoriser l□information contenue dans le Tableau 1 en suivant deux stratégies distinctes associées à deux types de traitements. A ce niveau, on peut distinguer les analyseurs de .Log classiques et l□analyse qui repose sur la construction de réseaux.

Nous nous proposons de faire une présentation successive de ces deux méthodes avant de voir l□intérêt de la méthode réseau par rapport aux analyses traditionnelles:

- Les techniques utilisées dans les analyseurs de .Log classiques:

Les analyseurs de .Log traditionnels présentent des résultats sous forme de tableaux statistiques qui décrivent par des comptages les différentes facettes du problème à représenter. Appliqués à un grand nombre de connexions, ces tableaux permettent de dégager une information de synthèse.

Le tableau 3 visualise, par exemple, les 30 pages du serveur du Crrm les plus consultées lors des 2869 connexions réalisées du 11/12/1996 au 20/12/1996. Ces pages sont classées par ordre de fréquence décroissante

Formes	Occurrence
0/entertainment/polar/pwhp.htm	402
0/vl/vlis.html	190
0/vl/metricts.html	138
0/entertainment/polar/polarweb/pwnovel.htm	124
0/vl/tech.html	119
0/entertainment/polar/polarweb/pwbooks.htm	117
0/entertainment/marseille/om/om.html	116
0/Welcome.html	97
0/entertainment/polar/polarweb/pwlinks.htm	93
0/auresys/present.htm	90
0/entertainment/polar/polarweb/pwtete.htm	78
0/entertainment/polar/polarweb/bilipohp.htm	77
0/orps/orps-home.html	76
0/entertainment/polar/recherch.htm	76
0/entertainment/polar/polarweb/pwinfos.htm	76
0/entertainment/marseille/marseille.html	75
0/entertainment/polar/polarweb/pwcritik.htm	73
0/entertainment/marseille/om/photos2/om-caen.html	71
0/entertainment/marseille/om/photos/photos.html	65
0/orps/u3-jerome/sommaire.html	62
0/entertainment/polar/instant.htm	62
0/entertainment/polar/813hp.htm	62
0/entertainment/entertainment.html	60
0/entertainment/marseille/highlight/highlight.html	57
0/projet_vrml/paca2.htm	56
0/entertainment/polar/813/813revue.htm	53
0/vrs_fr/garde.htm	52
0/softshell/betaprogram/beta.html	52
0/issi/issi-home.html	51
0/entertainment/polar/polarweb/pwexpres.htm	50

Tableau 3: Pages du serveur du Crrm les plus consultées

Le tableau 3 a été obtenu sous le logiciel Dataview (Rostaing, 1993) développé au Crrm mais de nombreux analyseurs de .Log génèrent ce genre de résultats.
Ce tableau est intelligible car les pages possèdent des noms signifiants.
On peut obtenir selon un processus analogue d□autres tableaux de synthèses qui exprimeront en fonction des besoins: la répartition du temps passé sur chaque page du serveur, la profondeur moyenne des utilisateurs du site etc.
Ce type d□analyse peut être complétée par l□analyse en terme de réseau que nous proposons d□introduire.

- Les techniques utilisées dans l'analyse en terme de réseaux:

Il est possible de construire différents types de réseaux à partir des données figurant dans le fichier reformaté. La caractéristique commune de ces analyses est d□obéir à une logique relationnelle.
Nous allons illustrer la démarche à partir du réseau le plus élémentaire qui puisse être construit. Il consiste à retenir du tableau 1 l□information correspondant au champ «relations ». On obtient les données du tableau 4:

```
relations:ND
relations:/cgi-bin/Count.cgi?tr=N&dd=C|df=polar.dat @/entertainment/polar/polarweb/pwtete.htm
relations:/entertainment/polar/polarweb/pwtete.htm @/entertainment/polar/polarweb/album.htm
relations:/entertainment/polar/polarweb/album.htm @/entertainment/polar/polarweb/pwcritik.htm
relations:/entertainment/polar/polarweb/pwcritik.htm @/entertainment/polar/polarweb/pwlinks.htm
relations:ND
relations:/vl/vlis.html @/vl/metricts.html
```

Tableau 4:L'extraction du champ « relations »

Cette information doit être transformée sous forme matricielle avant de pouvoir être représentée sous forme de réseau.
La transformation matricielle consiste à construire une matrice carrée symétrique. Dans la première colonne de cette matrice figure l'ensemble des pages consultées. Il en va de même de la première ligne. L'intersection entre la ligne i et la colonne j comporte un entier naturel qui correspond au nombre total de connexions lors desquelles ces deux pages ont été visualisées successivement.
Ce type de matrice peut être obtenue en utilisant le logiciel Dataview, produit par le Crrm.
La matrice associée au tableau 4 est présentée tableau 5:

	ND	/ENTERTAINMENT/POLAR/POLARWEB/PWTETE.HTM	/ENTERTAINMENT/POLAR/POLARWEB/PWCRITIK.H	/ENTERTAINMENT/POLAR/POLARWEB/ALBUM.HTM	/VL/VLIS.HTML	/VL/METRICTS.HTML	/ENTERTAINMENT/POLAR/POLARWEB/PWLINKS.HT	/CGI-BIN/COUNT.CGI?TR=N&DD=C|DF=POLAR.DA
nd	2	0	0	0	0	0	0	0
/entertainment/polar/polarweb/pwtete.htm	0	2	0	1	0	0	0	1
/entertainment/polar/polarweb/pwcritik.h	0	0	2	1	0	0	1	0
/entertainment/polar/polarweb/album.htm	0	1	1	2	0	0	0	0
/vl/vlis.html	0	0	0	0	1	1	0	0
/vl/metricts.html	0	0	0	0	1	1	0	0
/entertainment/polar/polarweb/pwlinks.ht	0	0	1	0	0	0	1	0
/cgi-bin/count.cgi?tr=n&dd=c|df=polar.da	0	1	0	0	0	0	0	1

Cette forme est présente deux fois dans le corpus

Ces deux formes sont associées dans une seule référence

Tableau 5: Matrice carrée symétrique

A partir de cette matrice carrée symétrique, il est possible de construire, à l'issu d'un processus automatique réalisé sous le logiciel Matrisme (Boutin et alii, 1995), un graphe appelé réseau. Ce graphe est composé de sommets et d'arcs. Les sommets correspondent aux différentes pages visualisées par les visiteurs. Un arc entre deux sommets signifie qu'un utilisateur au moins a utilisé le lien hypertexte existant entre les deux pages. Si on construit le réseau correspondant à l'exemple simple que nous avons considéré, on obtient le résultat présenté figure 3.

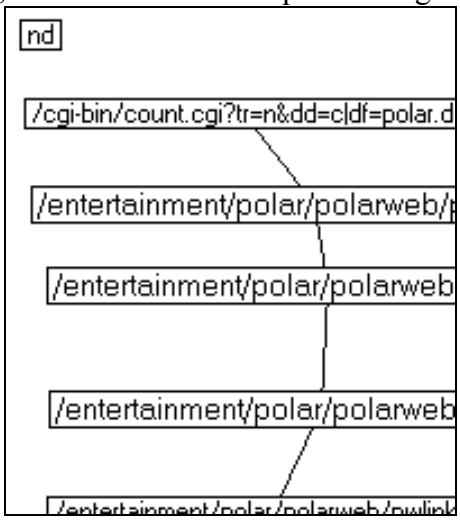

Figure 3: un exemple de réseau

Ce résultat appelle quelques commentaires.
- Il est d□une interprétation extrêmement intuitive. Le réseau de la figure 4 fait apparaître 2 groupes indépendants correspondant chacun à une connexion réalisée dans le mini exemple. La troisième connexion n□est pas affichée dans la mesure où l□utilisateur s□est déconnecté dès son arrivée sur le site. Cette forme de visualisation présente l□avantage de la clarté car les enchaînements entre les différentes pages sont clairement visualisés.
- Dans un tel réseau, la position des boîtes les unes par rapport aux autres n□a pas de sens. Les boîtes ont été disposées de manière à limiter les chevauchements entre les arcs du réseau.
- Les arcs ne sont pas signés si bien que lors d□une connexion particulière, il est impossible de déterminer à partir de l□observation du réseau quel a été le point de départ et le point d□arrivée de l□utilisateur sur le site.
- Lorsqu□on cherche à représenter sous forme de réseau un grand nombre de connexions, il n□est plus possible d□arriver à identifier les connexions individuelles dans la mesure où certaines pages ont été visualisées par plusieurs utilisateurs. Ces pages, lieux de passage obligé constituent les noeuds du réseau.

- La comparaison des méthodes:

Ces deux méthodes ne doivent pas être considérées comme concurrentes mais complémentaires l□une l□autre. Nous aimerions montrer ici le supplément d□information apporté par l□approche réseau. Lorsqu□un client se connecte sur un site, les pages qu□il visualise sont porteuses de sens mais on peut également s□intéresser à l□ordre dans lequel ces pages sont visualisées. Cet ordre est pris en compte par les liens retenus par le client. Les indicateurs traditionnels ne prennent pas en compte 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 par laquelle l□utilisateur a abordé le serveur.
Considérons un exemple pour illustrer ce propos. 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.

II- VALIDATION DE LA DEMARCHE: L'AUDIT DU SERVEUR DU CRRM

Notre objectif est ici de montrer la diversité des analyses qui peuvent être mises en □uvre en utilisant l□approche réseau.
Notre approche considérera successivement deux angles complémentaires du problème. Dans un premier temps, nous raisonnerons sur le réseau global. Ce réseau s□intéresse à toutes les connexions établies, quelles que soient leur durée, leur profondeur. Dans un second temps, nous examinerons quel est le parcours des visiteurs du site. Dans un troisième temps nous privilégierons l□analyse des visiteurs qui se sont connectés plus de 10 minutes.
Pour pouvoir analyser le mieux possible les réseaux présentés dans la suite de ce document, il nous a semblé judicieux d□introduire dans le tableau 6 les différents axes de recherche du site duCrrm, en quelques lignes chacun. A chacun de ces axes est associé un nom générique de page que l□on retrouve dans les différents réseaux suivants.

/vl/metricts.html
Librairie Virtuelle en sciences de l□information animée par LucQuoniam
/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 présnt à chacun de ces colloque

/orps/orps-home.html

L'Observatoire Régional de la Production Scientifique Provence - Alpes - Côte d'Azur est réalisé par le CRRM avec la collaborations 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 recupérées 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 crées, 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.

Tableau 6: Les axes de recherche présents sur le site du Crrm.

A- ANALYSE DU RESEAU GLOBAL:

L'exemple simple qui a été utilisé pour présenter la méthode et qui a trouvé son aboutissement dans le réseau de la figure 4 est trompeur. En effet, il donne l'illusion que la représentation sous forme de réseau est toujours extrêmement visuelle. Si, au lieu de considérer 3 connexions comme dans l'exemple précédent, nous nous intéressons aux 2869 connexions réalisées sur le site duCrrm pendant la période considérée, la visualisation du résultat devient confuse et le graphe parfaitement inextricable comme l'illustre le réseau présenté figure 4.

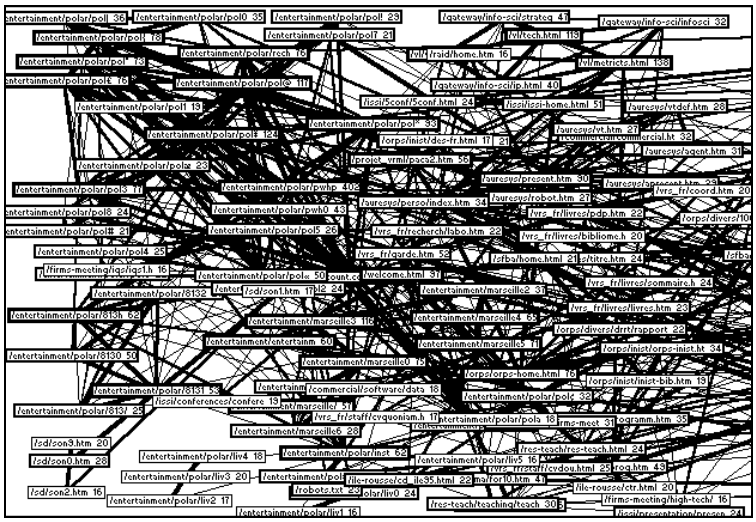

Figure 4: Réseau non filtré résultant des 2869 connexions.

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.

Un certain nombre d'analyses peuvent être conduites pour dégager de ce réseau des informations pertinentes. Ces analyses utilisent toutes la technique du filtrage. Filtrer le réseau va consister, selon le

cas, à supprimer certains sommets ou certains liens ou les deux à la fois. Nous allons illustrer la démarche de filtrage en considérant le filtre sur les paires.

Le point de départ de cette analyse consiste à remarquer que le réseau de la figure 4 est inextricable, du fait du trop grand nombre de liens qui y figurent. Or ces liens n□ont pas tous le même poids. Les liens les plus forts, correspondant aux arcs les plus épais, sont associés à des enchaînements de pages qui sont plus utilisés que d□autres. Si nous souhaitons nous intéresser aux liens hypertextes les plus souvent utilisés par les visiteurs, nous allons appliquer un filtre qui éliminera les liens les plus ténus. C□est ce qui a été fait dans la figure 5 où, pour faciliter la lisibilité du réseau, les liens de fréquence inférieure à 6 ont été supprimés.

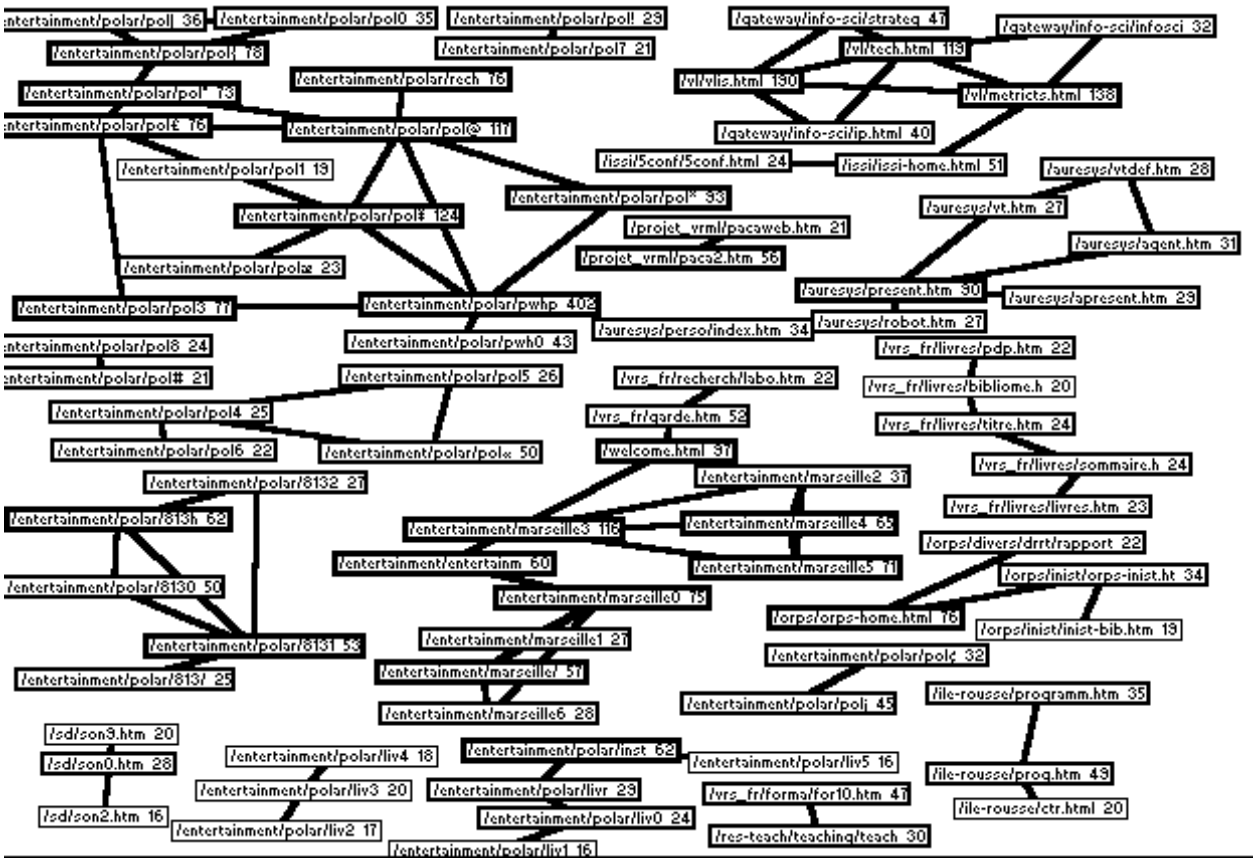

Figure 5: Réseau obtenu en retenant les paires supérieures ou égales à 6.

Dans chaque boîte figure le nom de la page suivi du nombre de clients différents qui l□ont visitée. Ce réseau fait apparaître clairement deux activités sur ce serveur: une activité ludique représentée sur la partie gauche du réseau et une activité recherche.

Une nuance doit être introduite pour rendre compte de l□importance de la partie ludique, paradoxale sur un serveur consacré à la recherche. On peut schématiquement distinguer deux grands types de structuration de pages sur un serveur. Le premier type renvoie à une organisation linéaire

L□autre type de structuration renvoie à une organisation en étoile. Dans un tel système, chaque page renvoie à de nombreuses pages potentielles si bien que l□utilisateur a un grand nombre de possibilités à chaque niveau.

Pour un même nombre x d□individus connectés à la page 1, le nombre d□individus qui se connecteront à la page 2 sera beaucoup plus élevé dans le premier cas que dans le second ceteris paribus. Si on effectue un filtre en supprimant les liens les plus faibles, on 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.
On en conclut donc que le choix d□un filtre sur les paires n□est pas neutre sur la structure du réseau.

B- ANALYSE DE PARCOURS:

On peut illustrer le parcours d□un visiteur sur le site du CRRM par la figure 6.

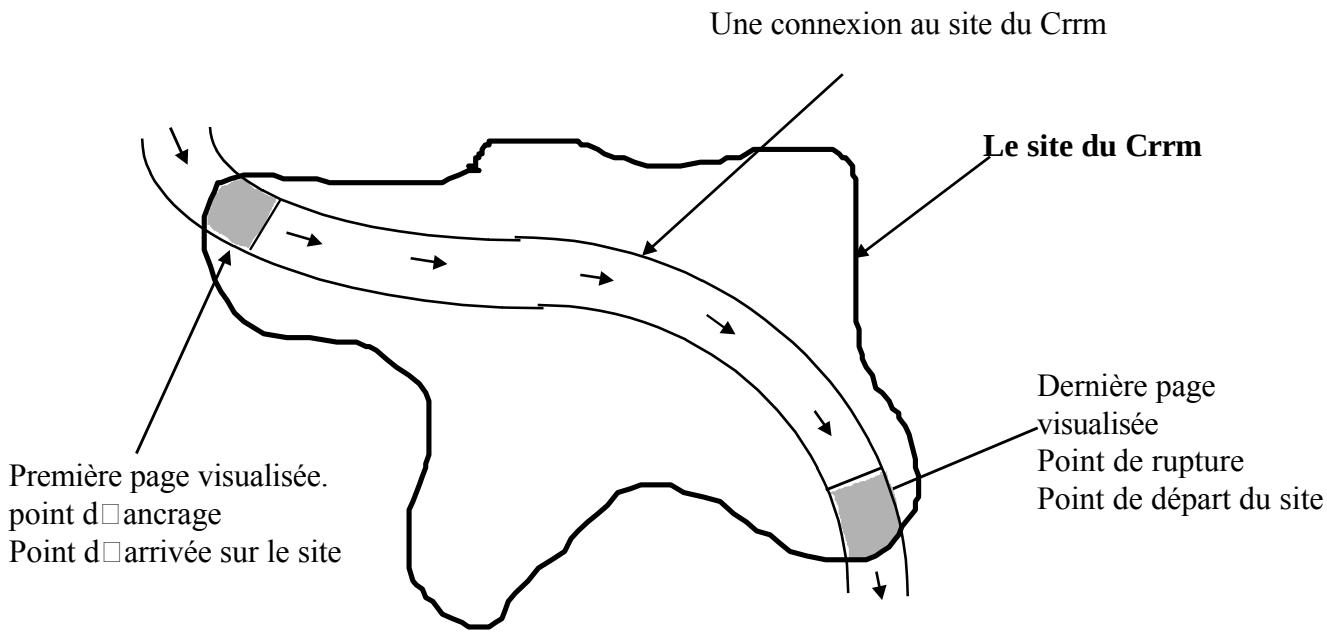

Figure 6: Parcours d'un utilisateur sur le site du Crrm

Deux moments privilégiés peuvent être objet d□analyse. Le premier s□intéresse à la première page que visualise le visiteur lors de sa connexion. Le second s□intéresse à la page sur laquelle le visiteur va quitter le site.
Nous avons choisi de privilégier le premier élément.

- Le début de la connexion:

Le graphe présenté figure 7 correspond au réseau construit à partir des deux premières pages visualisées par chaque visiteur. Pour ne pas alourdir la présentation de ce graphe, nous n□avons retenu que les paires supérieure ou égales à trois. Dans chaque boîte figure le nombre de fois où la page a été visualisée en premier lieu.
Plusieurs réflexions ressortent de l□observation de ce réseau.

* On peut remarquer la part relativement négligeable occupée par la page « welcome.html » qui n□a été utilisée que 7 fois en tant que page d□accueil par les visiteurs ce qui représente 0.5% des connexions réalisées. Ceci est surprenant car « welcome.html » est la page d□accueil du Crrm. Le site du Crrm est structuré autour de cette page qui a une double fonction. La première est de servir d□« aiguillage ». La seconde est de respecter une démarche pédagogique de présentation. Cette constatation du désintérêt pour la page prédéfinie par les développeurs comme étant la page d□accueil

du site peut animer deux réflexions, la première s□intéressant aux causes de ce phénomène, la seconde, plus fondamentale, s□intéressant à ces implications.

Deux raisons peuvent être invoquées pour rendre compte du faible attrait de la page welcome.html. La première est que les visiteurs du site du Crrm n□arrivent bien souvent pas sur ce site pour la première fois. Ils ont parfois eu l□occasion de s□y connecter par le passé et de faire un certain nombre de points de repaires sur les pages qui les intéressent. Ces pages, recueillies dans un répertoire d□adresses, sont accessibles directement lors d□une connexion ultérieure.

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. A titre d□illustration, la page «vlis.htm » observée 167 fois en première position est 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 duCrrm, une structure émergente informelle qui tire sa légitimité de son utilisation. L□existence de cette structure informelle et le faible chevauchement entre cette structure informelle et la structure formelle doit être soulignée.

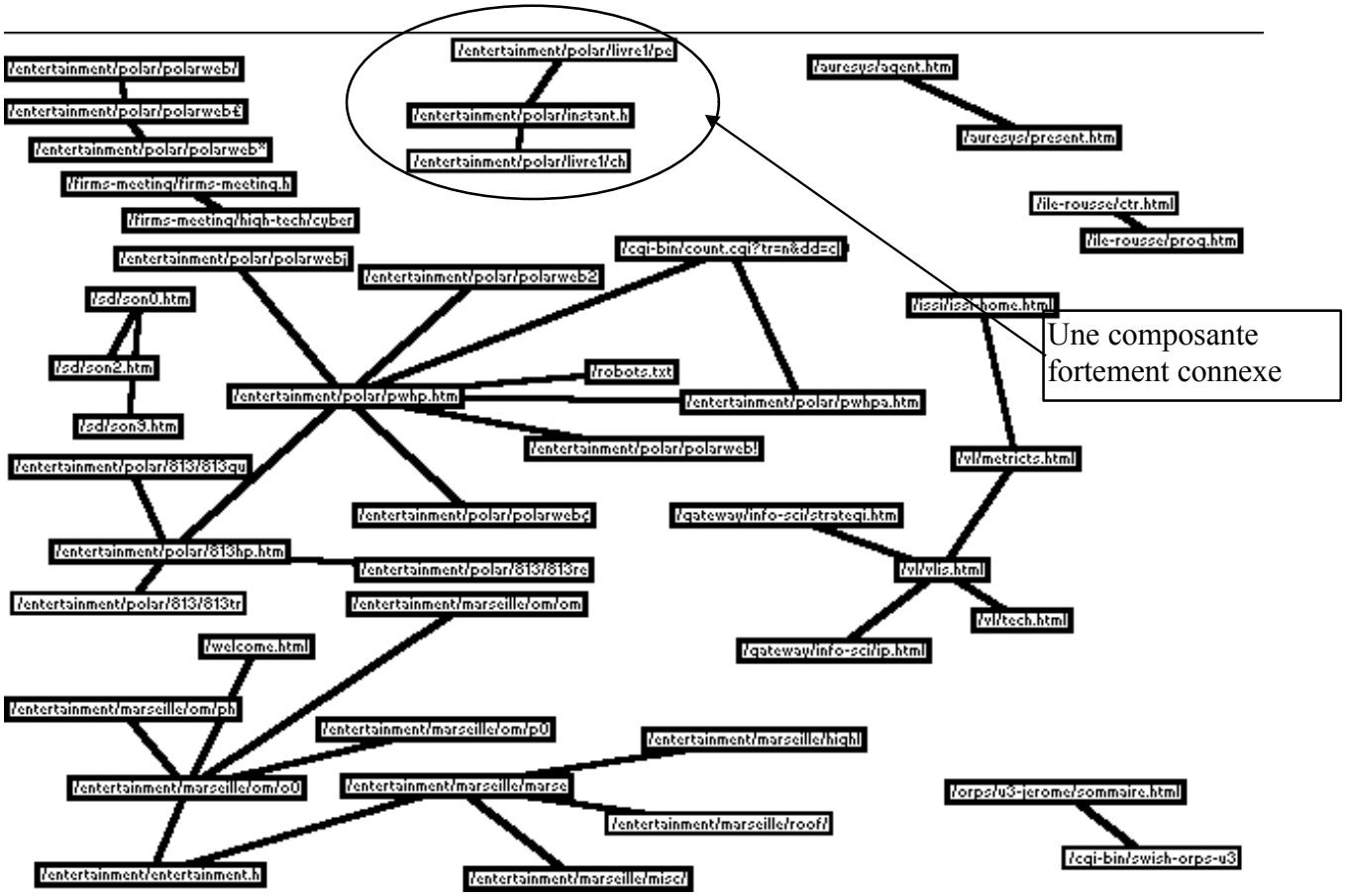

Figure 7:Réseau construit à partir du premier lien de chaque visiteur

* Ce réseau permet d□opérer des regroupements entre les différentes pages regroupées ici en 8 composantes fortement connexes. La partie ludique est la plus représentée sur ce réseau. Elle se subdivise elle même en plusieurs composantes étanches: Les deux principales sont formées à partir des deux racines « entertainment\polar » et « entertainment\marseille ».

C- ANALYSE SPECIFIQUE DES VISITEURS CONNECTES PLUS DE 10 MINUTES :

Nous allons nous intéresser aux visiteurs du site du Crrm en fonction de la durée de leur connexion. Nous avons choisi de privilégier les 169 visiteurs qui se sont connectés plus de 10 minutes. Ce seuil de 10 minutes a été arbitrairement choisi. 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, ceteris paribus, celles qui manifestent le plus d□intérêt pour le site. Elles méritent qu□on s□intéresse à elles de façon prioritaires.

L□analyse des 167 visiteurs du site duCrrm qui se sont connectés plus de 10 minutes a été appréhendée en retenant les fréquences de paires supérieures ou égales à 4.

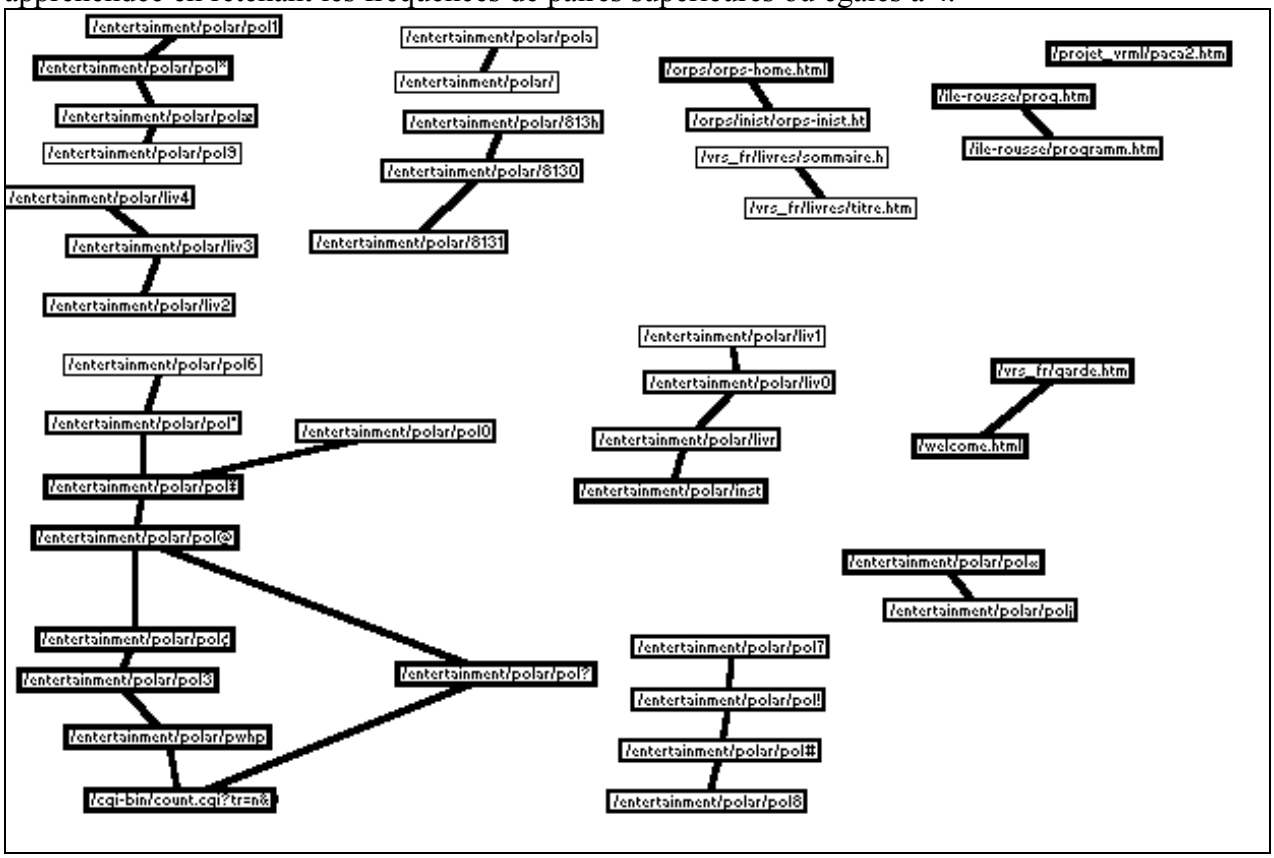

Figure 8: Réseau des utilisateurs restés plus de 10 minutes sur le site du Crrm en retenant les fréquences de paires supérieures ou égales à 4.

La figure 8 ne fait que confirmer les intuitions précédantes. La présence des pages ludiques sur ce réseau est forte chez les visiteurs qui se connectent plus de 10 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.

Pour conclure sur ce point, on doit certainement considérer que l□activité ludique de ce serveur n□est pas seulement une activité complémentaire de la partie recherche. Cette partie du site consacrée au polar est à elle seule un point d□attraction du site. Ceci est confirmé par la figure 11 qui indique que 243 visiteurs du site du Crrm ont, comme point d□entrée sur le site, la page «pwhp.htm » consacrée aux romans policiers.

Conclusion :

L□objectif de ce document est de présenter une méthode originale de traitement des fichiersLog. Cette méthode complète les analyses traditionnelles en apportant une dimension visuelle et en rendant possible une analyse de parcours ce que les autres analyses ne permettaient pas. En terme de recherche, cette analyse émergente offre trois perspectives intéressantes (Boutin, 1997) :

- La première repose sur la construction d□une méthodologie d□analyse systématique d□un site internet à partir de l□outil réseau.

- La seconde considère qu□un travail d□audit ne doit pas être réalisé uniquement à partir du fichier Log mais que l□interface avec le site lui même doit être privilégiée.
- La troisième perspective consiste à identifier des segments de visiteurs qui se caractériseraient par des comportements particuliers et qui pourraient constituer les cibles d□une démarche marketing.

Eléments de bibliographie :

- H. Rostaing, *Veille technologique et bibliométrie : concepts outils et applications*, thèse : Aix-Marseille III, 353 p, 1993.
- A. Degrenne, M. Forsé, *Les réseaux sociaux*, Editions Armand Colin, 1994
- E. Boutin, L. Quoniam, H. Rostaing, H. Dou, (1995*), A new approach to display real co-authorship and co-topicship through network mapping*, Acte du colloque « Fifth international conference on scientometrics & infometrics » Chicago, 7-10 Juin 1995.
- Eric Boutin, *Analyse du Log de la technopôle de l'Arbois*, WP, Laboratoire Le Pont, Mars 1997.