Application de l'analyse réseau à la modélisation de la visite d'un site web

J.M. Ferrandi, Eric Boutin

▶ **To cite this version:**

J.M. Ferrandi, Eric Boutin. Application de l'analyse réseau à la modélisation de la visite d'un site web. Recherche et Applications en Marketing (French Edition), 2001, 16 (3), pp.79-94. sic_00828606

HAL Id: sic_00828606

https://archivesic.ccsd.cnrs.fr/sic_00828606

Submitted on 31 May 2013

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Recherche et Applications en Marketing
http://ram.sagepub.com/

Application de l'analyse réseau à la modélisation de la visite d'un site web

Jean-Marc Ferrandi and Eric Boutin

Recherche et Applications en Marketing 2001 16: 79
DOI: 10.1177/076737010101600306

The online version of this article can be found at:
http://ram.sagepub.com/content/16/3/79

Published by:
SAGE
http://www.sagepublications.com

On behalf of:
afm
Association Française du Marketing

Additional services and information for *Recherche et Applications en Marketing* can be found at:

Email Alerts: http://ram.sagepub.com/cgi/alerts

Subscriptions: http://ram.sagepub.com/subscriptions

Reprints: http://www.sagepub.com/journalsReprints.nav

Permissions: http://www.sagepub.com/journalsPermissions.nav

Citations: http://ram.sagepub.com/content/16/3/79.refs.html

>> Version of Record - Sep 1, 2001

What is This?

Downloaded from ram.sagepub.com at UNIVERSITE DU SUD TOULON VAR on May 31, 2013

Recherche et Applications en Marketing, vol. 16, n° 3/2001

FENÊTRE SUR

Application de l'analyse réseau à la modélisation de la visite d'un site *web*

Jean-Marc Ferrandi *

Maître de conférences, LATEC, Université de Bourgogne, IUT de Dijon, Auxerre

Eric Boutin **

Maître de conférences, Laboratoire Le Pont, Université de Toulon et du Var, IUT de Toulon

RÉSUMÉ

L'objectif de notre recherche est de présenter une nouvelle méthode d'analyse du trafic sur un site *web* et de valider un nouveau modèle de la répartition du trafic entre les pages du site. Cette modélisation est fondée sur l'analyse réseau et les chaînes de Markov. La première partie présente les différentes approches disponibles pour mesurer l'audience et le trafic sur un site *web*. Après avoir exposé la méthodologie mise en œuvre, nous appliquons la modélisation envisagée au site Intranet de la SNCF et évaluons sa qualité en tant qu'outil de simulation. Enfin, nous développons les implications théoriques et managériales de notre recherche et exposons les voies de recherche futures.

Mots clés : Audience, trafic, fichiers *log*, analyse des réseaux sociaux, *web*, modélisation.

De plus en plus d'entreprises ou d'organisations fournissent à leurs clients, à leurs salariés et aux internautes qui le désirent des informations institutionnelles, opérationnelles ou commerciales sur leur site Intranet et/ou Internet. Chacune veut avoir une mesure des efforts qu'elle a entrepris sur le *web* et savoir si ses objectifs ont été atteints.

Contrairement aux médias traditionnels, le visiteur est à l'origine de la communication et est actif (Hoffman et Novak, 1996). Sa recherche part d'une démarche volontaire : demandeur d'informations, la durée de sa visite dépendra de la qualité de la réponse 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. Pour y arriver, le gestionnaire du site a besoin de connaître les raisons qui vont induire le passage sur son site et sur chacune de ses pages, soit en d'autres termes les éléments générateurs d'audience. Une transposition avec l'analyse conduite dans le domaine des moteurs de recherche s'avère intéressante pour comprendre les sources du trafic sur une page *web* donnée.

* Email : ferrandi@alcyone.u-bourgogne.fr.
** Email : boutin@univ-tln.fr
Les auteurs remercient les lecteurs pour leurs commentaires et Pierre Valette-Florence pour ses conseils.

Downloaded from ram.sagepub.com at UNIVERSITE DU SUD TOULON VAR on May 31, 2013

Le calcul d'indicateurs de pertinence de pages *web* est en effet couramment utilisé par les moteurs de recherche sur Internet afin de hiérarchiser les réponses à une requête d'un internaute. Les indicateurs de contenu apprécient la pertinence d'une page par son aptitude à rendre compte de la demande d'un internaute. Par contre, les indicateurs relationnels définissent l'audience d'une page par son positionnement au sein d'un ensemble d'autres pages. Ce type d'indicateurs est illustré par le « *PageRank* » adopté par le moteur de recherche Google (Brin et Page, 1998).

En transposant ces indicateurs de pertinence à la visite d'un site, nous pouvons formuler l'hypothèse que le trafic sur une page *web* dépend d'une part, de sa qualité intrinsèque, c'est-à-dire de la richesse et de la pertinence de l'information qu'elle apporte au visiteur, et d'autre part, de son positionnement ou de son degré de centralité par rapport aux autres pages du site (Drott, 1998). Nous privilégierons ici le second facteur et nous nous intéresserons à l'effet de la position de la page sur son trafic.

Le but de cette recherche est de montrer que le positionnement d'une page *web* dans son contexte est un critère qui permet de comprendre le trafic que connaît cette page. Pour ce faire, nous allons appliquer une analyse relationnelle, l'analyse réseau, au traitement d'un fichier texte, le fichier *log*, qui enregistre toutes les opérations effectuées par les visiteurs.

Notre recherche repose sur une analyse originale de ce fichier et débouche sur des mesures rétrospectives et prospectives du trafic sur un site. L'analyse rétrospective est fondée sur l'analyse des réseaux sociaux (Scott, 1991 ; Degenne et Forsé, 1994 ; Wasserman et Faust, 1994) utilisée dans d'autres cadres de recherche en marketing (Iacobucci, 1996 ; Boutin, Ferrandi, Valette-Florence, 1997). L'analyse prospective repose sur une modélisation des flux de passage entre les pages du site prises deux à deux. Cette modélisation est réalisée en appliquant au parcours des visiteurs des techniques de calcul utilisées en recherche opérationnelle (Isori *et alii*, 1988) et en théorie des graphes (Winne *et alii*, 1994).

Dans un premier point, après avoir présenté les approches développées par les praticiens et les chercheurs pour mesurer l'audience et le trafic sur un site *web*, nous soulignerons l'intérêt de l'analyse réseau. Dans la seconde partie, nous exposerons la méthode mise en œuvre pour obtenir une photographie du site et du parcours de ses visiteurs, et pour juger de la pertinence d'une page à l'aide d'indicateurs de centralité, puis nous modéliserons la répartition du trafic entre les pages du site. Ensuite, nous appliquerons la modélisation envisagée au site Intranet du fret de la SNCF [1] et nous évaluerons sa qualité en tant qu'outil de simulation. Enfin, nous présenterons dans un troisième temps les implications théoriques et managériales de notre recherche. En conclusion, nous soulignerons les apports, les limites et les voies de recherche de notre travail.

MESURES DE L'AUDIENCE ET SPÉCIFICITÉS DE L'ANALYSE RÉSEAU

De nombreux outils sont disponibles pour mesurer la fréquentation des sites *web*. Si les mesures centrées sur l'utilisateur qualifient l'audience, celles centrées sur le site quantifient uniquement le trafic sur le site. Après avoir exposé les différentes mesures de l'audience sur un site *web* utilisées par les praticiens et par les chercheurs, nous soulignerons l'intérêt d'une approche en termes de réseau par rapport aux analyseurs présents sur le marché.

La mesure de l'audience sur un site Internet

Chandon (2001) établit une distinction entre mesure du trafic et mesure de l'audience d'un site *web*. On parlera de mesure de trafic pour désigner l'analyse des connexions établies sur un serveur lorsqu'on ne dispose pas d'information relative aux internautes qui les ont réalisées.

1. Voici quelques éléments permettant de cerner le site Intranet de la SNCF : nous avons identifié 968 adresses IP distinctes. 181 pages différentes ont été consultées plus d'une fois, 35 742 liens ont été visualisés et 5 208 sessions de travail ont été réalisées.
Nous avons choisi d'utiliser un fichier Intranet car il permettait de lever les limites inhérentes aux fichiers *log* des sites Internet. Grâce à la technologie des marqueurs, ces limites ont aujourd'hui disparu. Ce choix ne restreint donc pas la portée de notre étude.

Downloaded from ram.sagepub.com at UNIVERSITE DU SUD TOULON VAR on May 31, 2013

Les mesures en terme d'audience sont, pour leur part, centrées sur les visiteurs du site dont les caractéristiques socio-démographiques sont connues et les activités de connexion sont suivies au moyen d'un logiciel placé sur leur ordinateur.

En raison de la taille que peut atteindre le fichier *log* (plusieurs giga-octets par jour pour un portail comme Yahoo), les responsables des sites ont besoin d'outils de « *datamining* » ou de techniques statistiques performantes pour analyser les informations. Deux types de mesure de l'audience et du trafic sur Internet sont possibles (Costes, 2000 ; Hussherr et Rosanvallon, 2001). Les approches site « *centric* » mesurent le trafic alors que celles centrées sur l'utilisateur mesurent l'audience du site.

• **La mesure *centrée sur l'utilisateur*** se fonde sur le comportement de l'internaute et est développée par des sociétés de panel (Coffey, 2001 ; Lohse *et alii*, 1999). La fiabilité de leurs résultats dépend de la composition et de la taille de leurs panels. Cependant, les panelistes ne peuvent pas encore garantir la représentativité de leurs échantillons en raison de l'instabilité de la population internaute (Dreze, Kalyanam et Briggs, 2000).

• **La mesure *centrée sur le site*** s'appuie sur l'analyse d'un fichier texte, qui enregistre le parcours du visiteur soit sur le serveur même du site, soit sur d'autres sites *web* (Dreze et Zufryden, 1998). Dans le premier cas, l'analyse repose sur l'exploitation du fichier *log* ou fichier trace, dans le second sur celle de *cookies*. Le fichier *log* est couramment exploité pour classer les pages selon leur fréquence de visualisation, évaluer la navigation sur le site, déterminer les pages à renforcer ou à alléger, ou pour améliorer la compréhension du parcours des visiteurs (Stout, 1997).

Les analyseurs de fichiers *log* présents dans le commerce, tels que *WebandStats* [2], *E-Stats* [3] ou *Weboscope* [4], se situent en aval du fichier *log* et restituent une information de synthèse souvent sous la forme de tableaux statistiques. L'objectif est principalement d'indiquer aux décideurs le nombre, la provenance et le comportement des internautes. Ces analyseurs professionnels appréhendent le trafic à partir de la notion de session [5].

2. http://www.tableaudebord.fr/home.htm
3. http://www.estat.com
4. http://www.weborama.com

La mesure de l'audience sur Internet : un champ de recherche nécessaire

Il n'existe pas à l'heure actuelle de *consensus* et de standards pour analyser le trafic sur le *web* en raison des problèmes méthodologiques et techniques soulevés par les deux types de mesure (Lendrevie, 2000). Etant donnée la difficulté à extraire des données significatives pour améliorer la compréhension du comportement de navigation des internautes, variable fondamentale en marketing (Peterson *et alii*, 1997), les chercheurs ont tout d'abord concentré leur attention sur les moyens capables d'améliorer les mesures existantes (Chi *et alii*, 2000, Novak et Hoffman, 1997 ; Pirolli *et alii*, 1996).

Lee et Leckenby (1998) ont proposé d'opter pour l'index de trafic [6], un critère composite de la couverture du site, de la fréquence et de la durée de visite. De plus, en 1999, ces chercheurs ont montré l'impact de la période de mesure sur le classement du site et sur l'estimation de son trafic. Le trafic varie en effet selon l'heure et le jour (en semaine ou durant le *week-end*). Dreze, Kalyanam et Briggs (2000) ont pour leur part souligné l'intérêt de combiner les approches centrées sur le site et sur l'utilisateur au moyen d'un modèle bayésien. En outre, certains chercheurs se sont intéressés aux lois statistiques suivies par le comportement de navigation des internautes. Ainsi, le nombre de liens retenus par les visiteurs (Levene, Borges et Loizon, 2001), les visites des sites et la répétition des visites (Montgomery et Faloutos, 2000) suivent des lois zipfiennes [7] (Zipf, 1949).

5. Une session désigne l'ensemble des pages parcourues par un visiteur lors d'une visite sur le site.
6. Trafic Index = couverture x fréquence x durée.
7. 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. Si on représente les différentes modalités d'un champ en abscisse en les classant d'après leurs fréquences d'apparition décroissantes dans le *corpus*, on obtient la courbe de Zipf. Cette courbe traduit l'existence d'un petit nombre de modalités à faible fréquence. Ce type de distribution rend inopérante la théorie des moments : la moyenne n'a pas de sens véritable et la variance est infinie. Toutes les analyses reposant sur l'hypothèse de normalité de la loi sont inopérantes.

Downloaded from ram.sagepub.com at UNIVERSITE DU SUD TOULON VAR on May 31, 2013

La répétition des visites par les internautes sur les sites pouvant être révélatrice de leur fidélité, des recherches ont été menées soit pour mesurer son importance (Cockburn et McKenzie, 2001), soit pour modéliser ce comportement (Lee, Dreze et Zufryden, 2000). Enfin, différentes mesures ont été mises en place pour améliorer l'évaluation et l'efficacité du contenu promotionnel ou publicitaire (Dreze et Zufryden, 1997 ; Kalika et Bourliataux-Lajoinie, 2001) et le choix des sites *web* par les visiteurs d'un portail (Goldfarb, 2001).

Globalement, si les recherches conduites en sciences de l'information et en intelligence artificielle ont été orientées vers l'outil (le site), les chercheurs en marketing et en comportement du consommateur se sont intéressés à la mesure du trafic et de l'audience des sites *web* pour mieux appréhender les facteurs susceptibles de mener le visiteur à cliquer sur une bannière publicitaire ou à acheter.

Différentes analyses du parcours des visiteurs semblent prometteuses à Pitkow (1997) : les courbes d'usures (Pitkow et Kehoe, 1995), les chaînes de Markov (Guzdial, 1994) ou les réseaux sociaux. Les modèles de Markov (Pirolli et Pitkow, 1999) peuvent être appliqués au comportement de navigation pour considérer les parcours des visiteurs sur un site *web*, comme les portails (Lee, Zufryden et Dreze, 2000). Pour notre part, nous avons choisi de développer une approche réseau couplée à une modélisation des flux de passage entre les pages d'un site, qui relève de la théorie des chaînes de Markov, pour analyser et prévoir les modes de navigation à partir du fichier *log*. L'analyse réseau, recommandée par Chatterjee, Hoffman et Novak (1998), présente un certain nombre d'avantages par rapport aux analyseurs développés dans le commerce.

L'intérêt de l'analyse réseau

L'approche réseau apportera des informations supplémentaires à celles fournies par les analyseurs présents sur le marché. Quand un client se connecte sur un site, les pages visualisées sont porteuses de sens. Or, on peut aussi s'intéresser à leur ordre de visualisation. Cet ordre prend en compte les liens retenus par le visiteur. L'analyse réseau envisagera la dimension séquentielle de la consultation et cherchera à enrichir l'information statique par le sens donné aux liens, reconstituant ainsi la démarche de l'utilisateur.

La consultation d'une page par un grand nombre de visiteurs 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 descriptive menée par les principaux logiciels du marché ne permet pas d'apprécier le second critère. En revanche, une représentation sous forme de réseau permettra de visualiser cette page dans son contexte et de la caractériser par un certain niveau de centralité.

La caractéristique générale de l'approche réseau est de ne pas s'arrêter à la vision statique et statistique des choses proposée par les analyseurs de *log* traditionnels, mais de positionner les pages dans leur contexte. On passe ainsi d'une démarche purement descriptive à une démarche analytique qui peut ouvrir la voie à des recommandations prospectives.

ANALYSES RÉTROSPECTIVE ET PROSPECTIVE DU TRAFIC SUR UN SITE

Les analyses complémentaires que nous allons présenter se situent à deux niveaux différents. La première, fondée sur l'analyse réseau, est de type macro. Ayant pour référent le site, elle cherchera à donner une photographie du parcours des visiteurs sur le site et à définir la pertinence d'une page ou d'une thématique à l'aide d'indicateurs de centralité. Par contre, la seconde est une analyse micro. Elle déterminera les probabilités de passage des visiteurs entre chaque couple de pages du site et permettra de disposer d'un outil de simulation de l'impact d'une modification du site.

L'analyse rétrospective d'un site Internet au moyen de l'analyse réseau

Notre approche repose sur la construction d'un graphe appelé réseau. Les sommets de ce graphe sont les différentes pages du site analysé. Un arc entre deux sommets signifie qu'un visiteur au moins est

Downloaded from ram.sagepub.com at UNIVERSITE DU SUD TOULON VAR on May 31, 2013

passé d'une page à l'autre. L'interprétation des réseaux peut se faire de manière intuitive par l'observation visuelle du graphe, mais aussi en s'aidant d'indicateurs de synthèse et de centralité qui permettent de rationaliser l'analyse en extrayant un certain nombre de sommets aux propriétés particulières. Nous raisonnerons d'abord sur le réseau global, qui retranscrit toutes les connexions établies, quelles que soient leur durée ou leur profondeur, puis sur des réseaux particuliers pour améliorer notre compréhension du comportement séquentiel des visiteurs.

L'analyse du réseau global

Le réseau visualisant le parcours de l'ensemble des visiteurs d'un site sur la période considérée renvoie généralement à un graphe parfaitement inextricable. Le réseau, graphe fidèle à la réalité, ne fait en effet que retranscrire le réel avec le moins de déformation possible. Si la réalité est complexe, par corollaire, le réseau l'est aussi. Une opération de filtrage du réseau est alors nécessaire pour dégager des informations pertinentes qui pourront être présentées sous forme de cartographie à l'utilisateur.

La répartition du trafic entre les pages d'un site *web* obéit à deux grandes logiques en interaction. Si certaines pages sont plus attractives en raison de leur forme et/ou de leur contenu, d'autres sont plus visitées du fait de leur position privilégiée dans l'architecture du site. Leur fréquentation n'est donc plus uniquement associée à leur caractéristique intrinsèque mais à leur surexposition au flux de visiteurs. Afin de mettre en évidence cette surexposition, nous allons privilégier les filtres sur les connectivités et sur les paires [8].

• **Le filtrage des connectivités** [9] est fondé sur le fait que la complexité du réseau est souvent due à un petit nombre de sommets qui ont un grand nombre de relations avec beaucoup d'autres. Retenir uniquement les pages les plus liées aux autres permet de dégager la métastructure du réseau, les sommets les plus centraux au sens de la centralité de degré [10].

• **Le filtrage des paires** est mis en œuvre en raison du trop grand nombre de liens qui figurent dans le réseau initial. Or, ces liens n'ont pas tous le même poids. Les plus forts sont associés à des enchaînements de pages plus empruntés que d'autres. Ce filtre permet de représenter les pages du site reliées par les flux de passage les plus denses et d'identifier des isthmes ou points d'articulation du réseau. Pour prendre une analogie avec le réseau routier, on cherche à identifier une carte sur laquelle seraient représentées les villes (pages du site) qui sont situées sur les axes principaux de circulation (autoroutes, nationales) et qui connaissent de ce fait les flux de passage les plus importants.

Ces deux filtres peuvent être mis en œuvre automatiquement par le logiciel d'analyse de réseau *Matrisme* (Boutin, 1999). Toutefois, leur détermination automatique n'est possible que si la distribution statistique du champ étudié suit une loi zipfienne (Zipf, 1949). Lhen *et alii* (1995) ont montré que les distributions statistiques observant une telle loi peuvent se décomposer en trois parties à partir de la notion d'entropie de Renyi [11] : les pages triviales, peu nombreuses mais très fortement liées aux autres, les pages qualifiées de bruit, fort nombreuses mais peu reliées aux autres et les pages intéressantes qui sont reliées aux autres un nombre moyen de fois dans le cas d'un filtrage des connectivités. Si nous filtrons le réseau global en retenant les pages les plus connectées entre elles, nous obtenons une liste de pages qui vont recevoir du fait de leur positionnement central dans le réseau un flux de passage important. Leur attractivité forte et leur trafic dense sont indépendants de leur qualité intrinsèque.

8. Nous ne présenterons pas ici tous les filtres disponibles mais illustrerons le *processus* de filtrage en fonction de notre problématique.

9. On appelle connectivité associée à un sommet x, le nombre d'arcs partant de ce sommet.

10. 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.

11. 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*.

Downloaded from ram.sagepub.com at UNIVERSITE DU SUD TOULON VAR on May 31, 2013

Grâce à l'analyse réseau, il est donc possible de représenter des facettes complémentaires de la réalité complexe que nous cherchons à appréhender. 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.

La compréhension du comportement séquentiel des visiteurs

La construction de réseaux particuliers répond au besoin de savoir si le trafic sur le site est le même quels que soient le niveau d'analyse, la partie de la connexion et le type de visiteurs. Ces différents réseaux permettent d'affiner l'analyse et d'améliorer notre compréhension du comportement de visiteurs particuliers.

• **Le niveau d'analyse :** L'analyse peut être conduite au niveau d'un groupe de pages correspondant à un thème particulier ou d'une page spécifique. L'analyse réseau permet de réaliser un zoom pour comprendre la raison de la centralité du thème ou de la page considérée et du comportement de leurs visiteurs.

• **La partie de la connexion qui nous intéresse :** Les premiers et derniers liens réalisés par les visiteurs sont deux moments privilégiés de l'analyse. Les pages qui constituent des points d'entrée correspondent-elles aux pages qui ont été définies par les concepteurs du site comme étant les pages d'accueil ou existe-t-il une logique de navigation de l'internaute qui échappe aux schémas définis par le concepteur ?

• **La catégorie de visiteurs qui nous intéressent :** Tous les visiteurs du site ou certaines catégories d'entre eux peuvent être appréhendés en fonction du temps passé sur le site ou de tout autre critère comme leur origine géographique. Si le temps passé sur le site est un indicateur du degré d'intérêt manifesté par le visiteur, les personnes qui se sont connectées le plus longtemps, méritent que l'on s'intéresse à elles en priorité.

Dans les analyses que nous venons de présenter, le référent est le site. Cette démarche correspond aux besoins d'une analyse macro du site et met en évidence la centralité de certains thèmes ou *a contrario* l'existence de branches mortes. Ultérieurement, notre analyse se situera à un niveau micro. Notre objectif sera de définir, toujours sur la base du fichier *log*, une matrice comportant les probabilités de passage entre chaque couple de pages du site. Notre analyse aura ainsi une portée beaucoup plus fine et opérationnelle. Son intérêt est d'offrir au concepteur du site un outil pour rapprocher la page qu'il a construite du comportement des visiteurs à son égard. On pourrait fort bien rapprocher à ce niveau la probabilité de clic sur tel bouton plutôt que tel autre sur une page donnée avec des considérations ergonomiques pour mettre en évidence les facteurs sous-jacents dans le comportement du visiteur.

L'analyse réseau a permis de déterminer la pertinence des pages en fonction de leur centralité. La modélisation que nous allons maintenant présenter va aider le concepteur du site à améliorer le positionnement de chaque page en augmentant les flux de visiteurs en fonction des objectifs qu'il poursuit.

Modélisation de l'activité d'un site web

Nous allons envisager les différentes étapes nécessaires pour modéliser les flux de passage entre les pages du site prises deux à deux. Le but est de disposer d'un modèle capable de prédire l'évolution de la répartition du trafic entre les pages d'un site après sa modification indépendamment de toute information sur la qualité intrinsèque de chaque page du site. Ce modèle, construit à partir des données empiriques fournies par le fichier *log*, repose d'une part, sur la représentation matricielle du flux de passage entre les pages du site prises deux à deux et d'autre part, sur l'étude de la propagation d'une arrivée exogène dans le système. Il est présenté dans la fenêtre 1. Reposant sur des matrices de transition, il relève de la théorie des chaînes de Markov. Nous allons maintenant comparer le résultat du modèle proposé au résultat réel et évaluer sa capacité à estimer l'effet d'une modification du site.

Notation

Le fichier *log* garde la trace du cheminement des internautes sur le site *web* considéré. Une telle trace est définie par de nombreuses informations parmi lesquelles nous retiendrons :

1. Le nom de la page visualisée. Pour simplifier l'écriture, nous ferons correspondre à chacune des m pages du site *web* un numéro compris entre 1 et m

Downloaded from ram.sagepub.com at UNIVERSITE DU SUD TOULON VAR on May 31, 2013

2. La date et l'heure à laquelle cette page a été visualisée.

Pour les besoins de la modélisation, nous devons construire deux indicateurs nouveaux à partir des données contenues dans le fichier *log* :

3. La session de travail de l'internaute. On identifie p sessions
4. La profondeur d'une page qui correspond au nombre de liens hypertextes plus un sur lesquels le visiteur a cliqué avant d'arriver sur cette page.

Ainsi une trace est définie de façon univoque par quatre variables. (T,i,j,k,d) désignera le fait que la page dont le numéro est i a été visitée lors de la $j$ième visite avec un niveau de profondeur de k à la date d.

En utilisant cette notation, le fichier *log* peut être présenté schématiquement par p lignes, chacune décrivant le parcours réalisé lors d'une session.

La représentation du flux de passage entre pages

Considérons une période de temps Δ_t. Au début de cette période, le visiteur arrive sur une page du site. A son terme, trois événements peuvent se produire : soit le visiteur utilise un lien hypertexte pour naviguer vers une nouvelle page du site, soit il quitte le site, soit il demeure sur la page en question. L'intégration, dans le modèle, de l'ensemble des pages visualisées par un visiteur oblige à s'intéresser aux flux de passage d'une page à l'autre. Matérialiser les flux de passage entre les pages revient donc à s'intéresser à la page, si elle existe, qui va être visualisée à la date $d + \Delta t$ à l'issue la trace $T(x,j,k,d)$.

1. Lorsqu'il n'est pas possible de trouver une page y telle qu'il existe $T(y,j,k+1,d')$, l'internaute a terminé sa session sur la page x et a quitté le site.
2. Si à l'issue de la page $T(x,j,k,d)$, l'internaute a visualisé la page $T(y,j,k+1,d')$, mais si $(d'-d)$ est supérieur à une durée Δt fixée arbitrairement, nous considérerons que l'internaute est resté sur la page x.
3. Si $T(x,j,k+1,d')$ existe et $(d'-d)$ est inférieur à Δt, l'internaute est passé de la page x à la page y.

Pour permettre de définir les probabilités de passage entre chaque couple de page, nous devons agréger les résultats correspondant au passage des pages deux à deux et introduire pour cela les notations suivantes :

4. $Z(x)$ correspond au nombre de fois où une session se termine par la consultation de la page x,
5. $A(x)$ correspond au nombre de fois où une session commence par la consultation de la page x,
5. $B(x)$ correspond au nombre de fois où l'internaute est resté un temps supérieur à Δt sur la page x,
6. Quels que soient x et y variant de 1 à m, $N(x,y)$ désigne le nombre de fois où un visiteur est passé de la page x à la page y après avoir passé un temps inférieur à Δt sur la page x

$$N(x) = \sum_{i=1+m} N(x,i) \tag{1}$$

7. $F(x)$ désigne le nombre de fois où la page x est présente dans le *corpus*.

$$F(x) = Z(x) + B(x) + N(x) \tag{2}$$

Après avoir évalué ces trois probabilités, nous pouvons établir une matrice de passage E carrée de taille m représentée dans le tableau 1 dans laquelle la valeur située à l'intersection de la ligne x et la colonne y correspond à la probabilité de passer de la page x à la page y. La première diagonale de la matrice correspond à la probabilité de rester sur la page entre deux périodes.

$$E(x,y) = N(x,y)/F(x) \tag{3}$$

$$E(x,x) = B(x)/F(x) \tag{4}$$

Tableau 1. – Matrice de transition

page	1	2	...	i	...	m
1	E(1,1)	E(1,2)		E(1,i)		E(1,m)
2	E(2,1)	E(2,2)		E(2,i)		E(2,m)
...						
i	E(i,1)	E(i,2)		E(i,i)		E(i,m)
...						
m	E(m,1)	E(m,2)		E(m,i)		E(m,m)

A ce stade nous pouvons donner une photographie précise du trafic sur le site considéré lors d'une période de temps Δt. Le modèle récapitule les probabilités de passage des pages prises deux à deux. Un de ses intérêts majeurs est de voir comment une arrivée exogène de visiteurs sur le site va se propager à ses

Downloaded from ram.sagepub.com at UNIVERSITE DU SUD TOULON VAR on May 31, 2013

différentes pages. Pour mesurer l'impact d'une modification du site sur la répartition de son audience entre ses pages, il faut intégrer au modèle les flux d'entrée sur le site.

Propagation aux différentes pages d'une arrivée exogène sur le site

L'étude de la propagation d'une entrée exogène à l'ensemble du système s'effectue en utilisant une propriété du produit matriciel qui a été largement utilisée dans les modélisations de comptabilité nationale et dans les analyses prospectives (Leontief, 1986 ; Xia et Yingzhong, 1990). Avant d'analyser le mode de propagation, il faut définir une entrée dans le système.

L'arrivée exogène d'un certain nombre de visiteurs sur le site peut être représentée par un vecteur ligne comportant m colonnes, chaque colonne précisant le nombre de visiteurs arrivant sur chacune des m pages à l'instant t. $A(x)/p$ est la fréquence des internautes arrivant sur la page x estimée à partir du fichier *.log*. Nous proposons d'injecter dans le système un vecteur V_t proportionnel au vecteur composé des $A(i)/p$ pour i allant de 1 à m.

Si nous multiplions le vecteur d'entrée exogène par la matrice de passage E, nous obtenons un vecteur comportant la position de ces visiteurs un temps Δt après leur arrivée sur le site. On note :

$$V_{t+\Delta t} = V_t x E \tag{5}$$

Il est ainsi possible instant après instant de suivre le flux des visiteurs et leur cheminement sur le site. Compte tenu du départ de certains visiteurs du site, le flux engendré par une arrivée exogène converge vers 0 en suivant une suite géométrique décroissante.

Nous disposons donc d'un modèle qui permet de simuler l'impact d'une arrivée exogène de visiteurs sur la fréquentation des pages.

Application au site Intranet du fret de la SNCF

Notre travail repose sur l'analyse du fichier *log* du site Intranet du fret de la SNCF du mois de septembre 1999. Ce fichier, de type *log* référentiel (Costes, 1998a et b), enregistre les requêtes reçues par le serveur et indique pour chaque page, celle précédemment visualisée par l'internaute. Au départ, les différentes lignes du fichier 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 l'identifiant du visiteur (son adresse IP) et de la date de connexion si l'on souhaite utiliser l'information disponible. Puisqu'un même visiteur peut se connecter à plusieurs reprises sur le site, pour isoler chaque session de travail, il faut considérer qu'après un certain laps de temps une nouvelle connexion a commencé [12]. Le *processus* de formatage des données s'accompagne de l'élimination des données jugées non pertinentes pour l'analyse, comme chacune des images associées à une page.

En passant d'une information brute à une information formatée, nous avons accompli un premier travail de création de valeur ajoutée. Nous allons maintenant évaluer la qualité de l'ajustement opéré et la valeur prédictive de la modélisation présentée.

L'évaluation du modèle repose sur une démarche en deux temps. Tout d'abord, nous comparerons les résultats du modèle avec les données résultant de l'exploitation statistique du fichier *log* du site Intranet de la SNCF. Le modèle retranscrit-il la réalité avec une précision suffisante ? Ensuite, nous évaluerons la qualité du modèle en tant qu'outil de simulation. Est-il à même de simuler l'impact des transformations d'un site sur la répartition du trafic entre ses pages ?

La qualité de l'ajustement du modèle à la réalité

Pour évaluer le modèle, nous avons choisi de le comparer avec les données résultant de l'exploitation statistique du fichier *log* du site de la SNCF. Le tableau comportant le nombre de visiteurs ayant visité chacune des pages nous donne le référent, le tableau constitué du vecteur V_5 [13] présente le résultat du modèle.

Si notre modèle est pertinent, la distribution observée durant le mois de septembre ne doit pas être très éloignée, à un coefficient près, des résultats du

12. Nous avons respecté la règle en vigueur : un arrêt de plus de trente minutes correspond au démarrage d'une nouvelle session.
13. Il est tout à fait possible d'appliquer le modèle jusqu'à épuisement total des flux de passage. Toutefois, cette option ne nous a pas semblé réaliste dans la mesure où les salariés de la SNCF réalisent en moyenne cinq clics de souris sur leur site. Nous avons donc choisi d'arrêter la simulation au bout de cinq itérations.

Downloaded from ram.sagepub.com at UNIVERSITE DU SUD TOULON VAR on May 31, 2013

modèle. Pour comparer ces deux distributions, nous avons effectué un ajustement de type puissance entre ces deux variables.

Nous avons appliqué cette démarche aux 181 pages du site étudié. Le graphique de la figure 1 fournit la distribution jointe des deux variables exprimées en valeur logarithmique. Un modèle de type puissance renvoie à une relation de type $Y = BX^A$ où Y désigne les observations réelles et X les résultats de la simulation avec ici $A = 0.76$ et $B = 109.94$. Nous obtenons un coefficient de corrélation de 0.91 ce qui permet de confirmer la validité du modèle.

La qualité prédictive du modèle

Le modèle proposé apporte un éclairage nouveau dans la mesure où il permet de passer d'une analyse globale à une analyse micro. L'analyse en termes de flux introduit une grande différence avec l'analyse de *log* traditionnelle. En effet, pour définir le trafic sur chaque page, nous considérons que le nombre de visites associé à une page est tributaire des liens qui arrivent sur cette page et de leurs probabilités respectives. Pour qu'une page soit visitée et qu'elle connaisse un flux important de visiteurs, une ou plusieurs des conditions suivantes doivent être respectées : cette page doit être une page d'accueil du site, le temps passé par le visiteur sur cette page doit être élevé, la probabilité de quitter le site sur cette page doit être faible, de nombreuses pages du site renvoient à cette page.

Lorsque le gestionnaire du site souhaite augmenter la visibilité de certaines pages par des actions ponctuelles, comme le rajout de liens hypertextes ou le changement d'une partie de l'architecture, il a besoin de disposer d'outils pour mesurer l'impact de ses modifications sur la répartition du trafic entre les pages de son site. Le modèle proposé lui offre-t-il la possibilité de simuler plusieurs transformations de ce type et d'évaluer leur impact [14] ?

Disposant uniquement du fichier *log* du mois de septembre 1999, nous avons testé la qualité du modèle en tant qu'outil de simulation en analysant l'impact d'une modification du site effectuée le 2 septembre 1999 et non prise en compte dans nos calculs antérieurs. A cette date, la page, */communic/bulletin/adv/tarifs7.htm*, a été intégrée au site.

Pour évaluer la validité du modèle comme outil de simulation nous avons déterminé le flux probable de visiteurs sur cette page. Pour ce faire, un élément

14. Pour que la simulation soit pertinente, elle ne doit pas remettre en cause da manière radicale l'organisation du site telle qu'elle est intégrée dans le modèle.

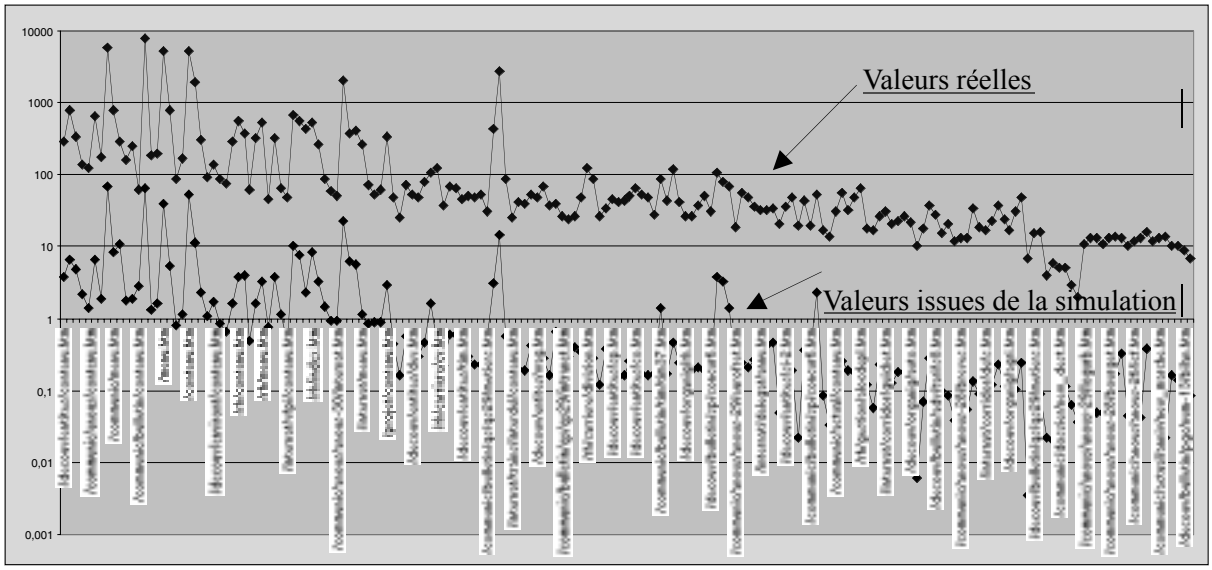

Figure 1. – Distribution jointe des deux variables exprimées en valeur logarithmique

Downloaded from ram.sagepub.com at UNIVERSITE DU SUD TOULON VAR on May 31, 2013

indispensable doit être pris en considération : la position de la page. Celle-ci est définie par deux paramètres : les pages qui pointent sur elle par un lien hypertexte et celles auxquelles elle renvoie. La volonté du gestionnaire du site de la SNCF a été de faire que le visiteur arrive sur cette page à partir de quatre pages. De plus, il a créé un lien unique entre elle et la page *communic/bulletin/contenu.htm*.

Afin de quantifier l'impact de la création de cette page en termes de fréquentation, nous devons estimer les probabilités de voir cette page référencée par les autres pages du site et les probabilités de passage de cette page aux autres pages. Pour estimer ces probabilités de passage, il serait utile de disposer du contenu des cinq pages. En effet, en dehors de toute considération ergonomique, il est difficile d'augurer de ces probabilités. Ne disposant pas d'une telle information, nous avons estimé ces différentes valeurs.

Considérons dans un premier temps la page *communic/contenu.htm* qui pointe sur la nouvelle page. Elle dispose de liens hypertextes vers d'autres pages avec les probabilités présentées dans le tableau 2.

Notre problème est de définir la probabilité de passage de la page *communic/contenu.htm* vers la page créée. Celle-ci se situe au troisième niveau de l'arborescence, puisque le nom de chaque page a été défini en respectant la structure hiérarchique du site. Nous avons estimé cette probabilité par la moyenne des probabilités des liens au départ de la page *communic/contenu.htm* qui appartiennent au même niveau de l'arborescence. Après calcul, nous observons qu'il existe une probabilité de 2.75 % pour que le visiteur passe de cette page à la page créée. Une fois cette probabilité estimée, toutes les probabilités au départ de cette page doivent être redéfinies pour faire en sorte que la somme des probabilités au départ de la page fasse bien 1. En réitérant ce calcul pour chaque page pointant vers la nouvelle page introduite, nous redéfinissons donc les probabilités de passage de l'ensemble de ces quatre pages.

Par contre, l'estimation des probabilités de passage au départ de la page *communic/bulletin/adv/tarifs7.htm* vers la page *communic/bulletin/contenu.htm* est plus délicate à définir puisqu'elle vient d'être créée. Il faut également définir pour cette page la probabilité d'y rester après un instant Δ_t et la probabilité de quitter le site après elle. Nous avons estimé ces valeurs à partir des valeurs des pages situées à un même niveau dans l'arborescence. Nous obtenons les probabilités schématisées dans la figure 2.

Tableau 2. – Les probabilités de passage de la page/communication/contenu.htm vers les autres pages du site de la SNCF

Pages	**/communic/contenu.htm**
/communic/anous/contenu.htm	4,44%
/communic/contenu.htm1	19,36%
/index.htm	7,27%
/communic/bulletin/contenu.htm	0,81%
/communic/anous/anous-30/format.htm	2,42%
/communic/anous/anous-30/edito.htm	3,03%
/communic/bulletin/qs/qs29/edito.htm	1,82%
/communic/anous/anous-30/technic.htm	6,66%
/communic/anous/anous-30/internat.htm	5,05%
/communic/dossier/sommaire.htm	3,03%
/communic/anous/anous-30/eurofret.htm	4,85%
/communic/anous/anous-30/decentr.html	6,46%
/communic/actual/anniv/sommaire.htm	2,22%
/communic/bulletin/tim/tim37.htm	1,01%
/decouv/bulletin/index.htm	6,66%
/communic/bulletin/sp/concur6.htm	1,82%
/communic/actual/contenu.htm	2,02%
/communic/mouvemt.htm1	16,96%
/communic/bulletin/pca/pca44.htm	0,40%
/communic/bulletin/pgc/num-13/bilan.htm	0,40%
/communic/bulletin/pca/pca43.htm	0,20%
/communic/bulletin/adv/tarifs5.htm	0,20%
/communic/actual/forum/sommaire.htm	1,41%
probabilité de quitter le site	1,49%
TOTAL	100,00%

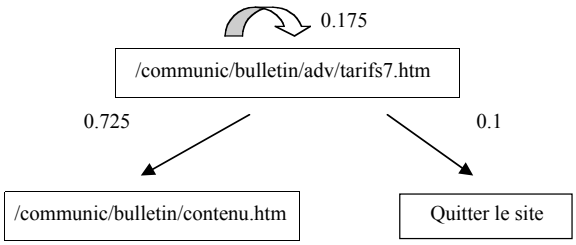

Figure 2. – Les probabilités relatives à la page /communic/bulletin/adv/tarifs7.htm

L'introduction de toutes ces probabilités dans la matrice de passage permet de lancer une nouvelle simulation. Le nombre de visites théoriques de la page nouvellement créée par simple application du modèle est alors de 1.36.

Downloaded from ram.sagepub.com at UNIVERSITE DU SUD TOULON VAR on May 31, 2013

Nous avons montré dans le point précédent que le modèle fournissait une estimation correcte de la réalité. L'équation de la droite d'ajustement puissance entre la théorie et la réalité est :

(1) $$Y = 109.94X^{0.76}.$$

Cette relation permet d'estimer le flux réel sur la page *communic/bulletin/adv/tarifs7.htm* en remplaçant X par 1.36. Nous obtenons une prévision de 138 visiteurs sur cette page. Si nous observons les résultats de la fréquentation de cette page pour le mois de septembre, nous obtenons une valeur réelle de 136 visiteurs, soit une erreur de 1.5 % par rapport à la valeur prévue. Cette valeur satisfaisante est trop ponctuelle pour nous permettre dès à présent de conclure à la validité du modèle en tant qu'outil de simulation.

LES IMPLICATIONS DES MESURES RÉTROSPECTIVE ET PROSPECTIVE DU TRAFIC D'UN SITE

Les analyses rétrospectives et prospectives du trafic sur un site présentées ont des implications tant sur les plans théorique que managérial. Ces implications portent sur l'efficacité des sites Internet, la segmentation des visiteurs, la compréhension de leur comportement de navigation et la fidélisation.

Les implications managériales

La mesure de l'audience et du trafic est d'un intérêt primordial pour les responsables marketing. Elle présente des enjeux économiques et stratégiques majeurs. Les informations relatives à la façon dont les individus accèdent à un site et y naviguent permettent en effet aux responsables marketing de comprendre le comportement des visiteurs sur leur site, de déterminer sa valeur financière, d'améliorer la notoriété de leur marque et de profiler les visiteurs afin de leur offrir des messages, des publicités et des contenus personnalisés.

L'analyse réseau permet de déterminer les points de passage obligés de l'utilisateur lorsqu'il se connecte sur un site, le parcours du visiteur type, l'articulation entre les différentes thématiques du site. Elle reconstitue la démarche par laquelle le visiteur a abordé et a quitté le site et prend en compte la dimension séquentielle de la consultation. 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 site. Cette amélioration de la compréhension du comportement des visiteurs est importante puisqu'elle peut aider à perfectionner le design du site dans le but d'économiser les efforts de l'utilisateur en le guidant directement vers les pages les plus pertinentes. Une amélioration du confort de navigation a en effet un impact positif sur le comportement des visiteurs et sur la mémorisation des messages (Chang, 2000 ; Ducoffe, 1996 ; Lynch et Ariely, 2000 ; Mandel et Johnson, 1999 ; Shneiderman, 1977). A cet égard, notre démarche permet également de simplifier la structure du réseau en éliminant les pages peu fréquentées ou ne jouant pas un rôle central, donc prépondérant, à l'intérieur du réseau. De même, notre approche offre la possibilité de mettre en évidence les pages les plus centrales et, par voie de conséquence, de déterminer celles qui doivent recevoir les éléments de communication et/ou publicitaires que l'administrateur du réseau désire mettre en avant.

Ainsi, l'utilisation de l'approche réseau devrait conduire à des procédures de segmentation qui aideraient à prendre des décisions relatives au positionnement et à la personnalisation des publicités. Segmenter les visiteurs à partir de leurs préférences et ensuite examiner les différences des réponses à la publicité pourrait permettre au concepteur du site d'identifier le contenu éditorial qui est le plus conducteur pour générer des clics et des visites répétées sur le site. Il est clair que l'approche développée ici, se situe à un niveau global pouvant masquer des segments sous-jacents avec des structures de réseau et donc des indices de centralité des pages différents. Modéliser les matrices de transition entre les pages visitées selon un *processus* de Markov au niveau latent permettrait de combiner les avantages premièrement décrits, c'est-à-dire les pages les plus importantes pour la construction du réseau, avec ceux

Downloaded from ram.sagepub.com at UNIVERSITE DU SUD TOULON VAR on May 31, 2013

d'une segmentation conduisant à l'identification de groupes de consommateurs ayant une pratique du site similaire. Mieux connaître les caractéristiques de ces segments donnerait alors la possibilité d'affiner et d'adapter le contenu des différentes pages du réseau.

Enfin, l'analyse micro est d'autant plus importante que se posent aujourd'hui les problèmes de savoir non seulement où positionner les bandeaux publicitaires ou les liens avec les sites partenaires pour qu'ils aient la meilleure efficience, mais aussi comment orienter le comportement des visiteurs pour qu'ils se dirigent comme souhaité. Les visiteurs d'une thématique particulière du site ont plus de chance d'être impliqués par une publicité se rapportant à ce sujet que la moyenne des internautes (Hussherr, 1999). Arriver à mieux connaître les segments constitutifs d'un réseau permettrait alors de mieux cerner et définir les publicités ou les bandeaux à insérer, passant ainsi d'une publicité générique à des publicités spécifiques, plus efficaces et surtout disposées à bon escient au suivi du réseau.

Les implications théoriques

De nombreux auteurs (Dreze et Zufryden, 1997 ; Ducoffe, 1996 ; Leckenby et Hong, 1998) ont souligné le besoin de mesurer l'efficacité des sites *web*. Celle-ci dépend du type de comportement adopté par le visiteur.

Les internautes peuvent présenter soit un comportement instrumental, orienté vers la réalisation d'un but comme la recherche d'informations, soit un comportement expérientiel (Hoffman et Novak, 1996). Ces comportements ont un effet sur le bénéfice utilitaire ou hédoniste attendu de la visite. De plus, les visiteurs ne se comportent pas de la même façon selon leur degré de contrôle de leurs actions sur Internet (Hoffman, Novak et Schlosser, 2000). Leur sensibilité au design et au contenu du site varie en outre selon leur degré d'implication (Cho, 1999). Malgré l'impossibilité de mesurer directement les raisons du comportement des visiteurs, l'analyse réseau présente l'intérêt d'appréhender le mode de navigation des utilisateurs et de déterminer si celui-ci est lié à un comportement instrumental ou expérientiel en fonction des centres d'intérêt révélés par l'utilisateur. Une analyse simultanée de l'approche réseau avec des mesures attitudinales devrait apporter des éclairages intéressants sur le mode de fonctionnement des internautes et sur leurs différences comportementales en fonction de facteurs situationnels ou personnels.

La relation entre le visiteur et l'entreprise se fait à travers le site avec lequel ils interagissent. L'évolution du comportement des utilisateurs dépend des réponses fournies. Internet permet des réponses adaptées grâce à son interactivité (Ghose et Dou, 1998). Or, une approche en termes de réseau est au cœur de l'interactivité. Le marketing interactif est en effet par essence fondé sur des réseaux. Il peut être étudié en utilisant les outils développés par l'analyse des réseaux sociaux pour aider à comprendre la signification des structures intriquées de réseaux.

Les deux approches proposées dans ce travail ont un intérêt indéniable en marketing. La navigation sur un site marchand s'apparente à la circulation dans un magasin. Dans ce dernier, l'entreprise a besoin de connaître les produits les plus demandés, d'améliorer leur agencement pour optimiser ses ventes. De même, sur son site, sa connaissance des pages les plus consultées, des liens les plus utilisés permet à l'entreprise d'aménager son magasin virtuel en fonction des attentes de ses visiteurs. Ces analyses permettent ainsi une extension du modèle de visite des magasins à un nouveau contexte : le marketing de circulation sur Internet. Guides pour le design et l'évaluation des sites *web*, ces approches sont un moyen de construire des relations à long terme avec les e-consommateurs.

La répétition des visites est une variable clé de la fidélité (Lee *et alii*, 2001). Elle augmente les chances d'acheter sur le site. Une tentation immédiate mais dangereuse est alors de cibler simplement les acheteurs fréquents plutôt que de suivre le comportement de navigation des visiteurs dans la durée, qui doit constituer la base de la segmentation des visiteurs (Moe et Fader, 2000). La modélisation, que nous avons présentée, aide à explorer les aspects dynamiques des *processus* de construction de la relation entre l'entreprise et le consommateur en analysant l'évolution du *processus* à la suite des modifications. Il est possible de mesurer si celles-ci ont un effet sur la fidélité au site.

Les implications théoriques et managériales de notre travail soulignent la nécessité de développer et d'approfondir cette recherche en intégrant une modélisation de la pertinence d'une page en fonction de la qualité de son contenu. Perkowitz et Etzion (2000) ou Kanawati et Malek (2000) ont développé les outils nécessaires pour créer des sites adaptatifs, c'est-à-

Downloaded from ram.sagepub.com at UNIVERSITE DU SUD TOULON VAR on May 31, 2013

dire un site qui s'améliore de lui-même à partir de l'apprentissage des modes d'accès des visiteurs sur la base de son fichier *log*. L'objectif est de rendre les pages populaires plus accessibles, de mettre en valeur les liens intéressants ou de connecter les pages visitées ensemble. Un tel développement des outils de mesure permettra réellement la mise en place d'un marketing interactif à un niveau micro avec des publicités adaptables selon le degré d'intérêt du visiteur pour une catégorie de produits par exemple.

CONCLUSION

Cette recherche avait comme objectif principal la présentation d'une nouvelle méthode d'analyse du trafic sur un site *web* et la validation d'un nouveau modèle de la répartition du trafic entre les pages du site. Bien qu'il s'agisse d'une première validation en matière de modélisation, les résultats s'avèrent très encourageants. Le modèle a notamment montré dans le cadre d'une analyse prospective du trafic sur un site de grande ampleur une très bonne fiabilité puisqu'il retranscrit 83 % de la réalité.

L'analyse réseau a permis de dépasser l'information fournie par les analyseurs de fichiers Log présents sur le marché et devrait offrir aux entreprises de nouveaux moyens pour mesurer le trafic sur leur site *web*. A la différence des approches traditionnelles qui raisonnent en termes de stocks et présentent des informations sous la forme de statistiques descriptives, le modèle présenté propose une mesure du trafic en termes de flux de passages entre les pages. Cette mesure a des incidences sur la perspective offerte. Au lieu de se contenter d'une mesure rétrospective du trafic, il est possible d'envisager une mesure prospective et d'évaluer les effets d'une modification du site sur la répartition de son trafic entre ses pages. L'intérêt d'une telle approche est de se situer non plus au niveau de l'ensemble du site mais à un niveau plus fin et opérationnel : celui des liens entre les pages et de leur effet sur la visite du site.

Notre approche est cependant limitée par la source d'informations utilisée. L'information contenue dans le fichier *log* doit être interprétée avec précaution en raison du mode de construction du fichier (Costes, 1998a et b). La mise en place de marqueurs sur chaque page permet aujourd'hui de contourner ces problèmes.

Diverses voies de recherche paraissent se profiler :

• Une première perspective de recherche future est de développer l'analyse micro. Grâce à son horizon temporel, cette analyse devrait aider le gestionnaire du site à estimer l'ampleur de la modification nécessaire pour obtenir un trafic *maximum* sur la page désirée.

• Une seconde voie de recherche est de développer des approches complémentaires entre l'analyse réseau du fichier *log* et le modèle de répartition du trafic. L'approfondissement de telles approches devrait apporter des réponses à la pertinence des choix de l'entreprise en matière de contenu, de convivialité, de facilité de navigation et d'interactivité.

• En outre, notre approche est purement instrumentale, ce qui constitue sa principale limite. Un développement conjoint avec des approches plus centrées sur le comportement réel des visiteurs semble primordial. La mise en place d'études quantitatives à l'aide de questionnaires permettrait d'évaluer de façon plus pertinente les motivations et les attitudes des individus. Si nous avons dans le cadre de cette recherche amélioré notre compréhension du comportement séquentiel des utilisateurs sur un site *web*, il ne nous est pas possible d'appréhender les fondements de leurs comportements. Il semble donc important de mettre en place des études qualitatives et notamment des plans expérimentaux en laboratoire pour améliorer notre connaissance du mode de fonctionnement de l'utilisateur devant un site commercial et pour saisir les raisons de son comportement (Burton et Walther, 2001).

• Par ailleurs, l'analyse réseau tant au niveau rétrospectif que prospectif a illustré le fait que le trafic que va connaître une page donnée est fonction de sa position plus ou moins centrale dans le réseau. Les pages d'un site connaissent en effet des flux de passage proportionnels à leur centralité dans le réseau. Cette remarque est sans doute de nature à émettre un jugement plutôt sceptique à l'égard de méthodes qui considèrent l'audience comme seul critère de mesure de la pertinence d'une page. Si le trafic d'une page n'est pas un indicateur pertinent pour rendre compte de la qualité d'une page pour un internaute, la durée de

Downloaded from ram.sagepub.com at UNIVERSITE DU SUD TOULON VAR on May 31, 2013

connexion sur une page est sans doute un indicateur plus approprié. Notamment, le temps de visite d'une page pourrait être un indicateur de sa capacité à retenir l'attention du visiteur et donc constituer un premier pas vers la persuasion (Harvey, 1997). Des mesures de l'attitude utilisées pour juger de l'efficacité de la publicité pourraient être transposées au niveau des pages pour juger de leur pertinence (Bruner et Kunar, 2000 ; Chen et Wells, 1999 ; Le Roux, 1998).

• Enfin, cette recherche utilise implicitement un modèle de Markov en ce qui concerne les matrices de transition d'une page à l'autre. Etendre la modélisation proposée aux modèles de Markov à un niveau latent (Poulsen et Valette-Florence, 1996 ; Bockenholt et Dillon, 2000) devrait permettre de mieux cerner la nature des flux entre les différentes pages visitées, mais aussi de classifier les visiteurs d'un site en groupes homogènes quant aux pages consultées et à leur manière de naviguer sur le site.

Compte tenu de l'importance économique que revêt la mesure d'audience sur le *web*, ce type d'application devrait, nous l'espérons, se développer à un rythme rapide. L'étude de la concurrence ne nous a pas permis, à ce jour, d'identifier des outils qui soient positionnés sur ces créneaux. Il s'agit donc pour l'instant d'une niche isolée appelée à un fort potentiel de développement.

RÉFÉRENCES BIBLIOGRAPHIQUES

Barwise P., Elberse A. et Hammond K. (2001), Marketing and the Internet: A research review, *Handbook of Marketing*, éds Weitz, Barton et Wensley, London, Sage, à paraître.

Bockenholt U. et Dillon W. (2000), Inferring latent brand dependencies, *Journal of Marketing Research*, 37, 1, 72-87.

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, Université des Sciences et des Techniques, Aix-Marseille III.

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.

Brin S. et Page L. (1998), The anatomy of a large-scale hypertextual web search engine, world wide web conference, 7, http://www7.scu.edu.au/programme/fullpapers/1921/com1921.htm

Bruner II G.C. et Kumar A. (2000), Web commercials and advertising hierarchy-of-effects, *Journal of Advertising Research*, 40, 1 /2, 35-42.

Burton M.C. et Walther J.B. (2001), The value of web log data in use-based design and testing, *Journal of Computer-Mediated Communication*, 6, 3, www.ascusc.org/jcmc/jcmcindex.html.

Chandon J.L. (2001), Elaboration du plan média, *La publicité: Théorie, acteurs et méthodes*, éd. E. Vernette, La Documentation Française.

Chang T.Z. (2000), Online shoppers' perceptions of quality of internet shopping experience, *Actes de la conférence Summer Educators*, American Marketing Association, 254-255.

Chatterjee P., Hoffman D.L. et Novak T.P. (1998), Modeling the clickstream: Implications for web-based advertising efforts, papier de recherche, http://www2000.ogsm.vanderbilt.edu/papers.html.

Chen Q. et Wells W.D. (1999), Attitude toward the site, *Journal of Advertising Research*. 39, 5, 27-37.

Chi E., Pirolli P. et Pitkow J. (2000), The scent of a site: A system for analyzing and predicting information scent, usage and usability of a web site, *Actes de la Conférence Human Factors in Computing Systems*, 161-168.

Cho C.H. (1999), How advertising works on www: Modified elaboration likelihood model, *Journal of Current Issues and Research in Advertising*, 21, 1, 33-50.

Cockburn A. et McKenzie B. (2001), What do web users do? An empirical analysis of web use, *International Journal of Human Computer Studies*, 54, 6, 903-922.

Coffey S. (2001), Internet audience measurement: A practitioner's view, *Journal of Interactive Advertising*, 1, 2, http://jiad.org/vol1/no2/coffey.

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

Costes Y. (1998b), La mesure d'audience sur Internet : un état des lieux, *Recherche et Applications en Marketing*¸ 13, 4, 53-67.

Costes Y. (2000), Comprendre et mesurer le profil et le comportement des internautes, *Revue Française du Marketing*, 177/178, 153-167.

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

Dreze X. et Zufryden F. (1997), Testing web site design and promotional content, *Journal of Advertising Research*, 37, 2, 77-91.

Dreze X. et Zufryden F. (1998), Is Internet advertising ready for prime time?, *Journal of Advertising Research*, 38, 3, 7-18.

Downloaded from ram.sagepub.com at UNIVERSITE DU SUD TOULON VAR on May 31, 2013

Dreze X., Kalyanam K. et Briggs R. (2000), Increasing panel data accuracy: An application to Internet panels, http://sbaxdm.usc.edu/Publications/list.html.

Drott M.C. (1998), Using web server log to improve site design, *Actes de l'International Conference on Systems Documentation*, 16, 43-50.

Ducoffe R.H. (1996), Advertising value and advertising on the web, *Journal of Advertising Research*, 36, 5, 21-35.

Ghose S. et Dou W. (1998), Interactive functions and their impacts on the appeal of Internet presence sites, *Journal of Advertising Research*, 38, 2, 29-43.

Goldfarb A. (2001), Analyzing website choice using click-stream data, papier de recherche, 103, http://e-commerce.mit.edu/cgi-bin/viewallpapers.

Guzdial M. (1993), Deriving software usage patterns from log files, rapport technique, #GIT-GVU-93-41, Graphics, Visualization and Usability Center, College of Computing, Georgia Institute of Technology.

Harvey B. (1997), The expanded ARF model: Bridge to the accountable advertising, *Journal of Advertising Research*, 37, 2, 11-20.

Hoffman D.L. et Novak T.P. (1996), Marketing in hyper-media computer-mediated environments: Conceptual foundations, *Journal of Marketing*, 60, 3, 50-68.

Hoffman D.L., Novak T.P. et Schlosser A. (2000), Consumer control in online environments, http://www2000.ogsm.vanderbilt.edu/papers.html.

Hussherr F.X (1999), La publicité sur Internet : un modèle économique dépendant de l'efficacité publicitaire, thèse de doctorat, Ecole Nationale Supérieure des Télécommunications, Paris.

Hussherr F.X. et Rosanvallon J. (2001), *E-communication. Tirer profit d'Internet : le sixième média...et plus encore*, Dunod.

Iacobucci D. (1996), *Networks in marketing*, Sage Publications.

Isori P., Carles D. et O'Brien G. (1988), Analyse matricielle des transitions (AMT), *Revue Education Médicale*, 11, 6.

Kalika M. et Bourliataux-Lajoinie S. (2001), L'analyse des comportements de navigation sur un site marchand, *Décisions Marketing*, 22, 79-86.

Kanawati R. et Malek M. (2000), COBRA : Une approche d'adaptation structurelle des sites web fondée sur une technique d'apprentissage à partir des traces d'accès utilisateurs et utilisant la méthodologie du raisonnement à partir de cas, présenté aux *Journées Internationales de Nîmes de Nouvelles Technologies de la Communication*, NTIC'2000, http://www.eerie.fr/~multimedia/nimestic/articles.html.

Le Roux A. (1998), L'attitude envers la publicité : facteurs explicatifs et rôle dans le processus de persuasion, papier de recherche n° 537, CEROG, Aix-en-Provence.

Leckenby J.D. et Hong J. (1998), Using reach/frequency for web media planning, *Journal of Advertising Research*, 38, 1, 7-20.

Lee S. et Leckenby J.D. (1998), An investigation of web-site ranking methods, présenté à l'Annual Meeting of the American Academy of Advertising, http://www.utexas.edu/coc/admedium/Ivory/3ASuckkee.html.

Lee S. et Leckenby J.D. (1999), Impact of measurement period on reach, frequency and ranking of web sites, *Journal of Current Issues and Research in Advertising*, 21, 1-10.

Lee S., Zufryden F. et Dreze X. (2000), A multivariate multinomial logit-Markov model and its application to consumer switching behavior on Internet, papier de recherche, Université Southern California, Los Angeles, CA. (http://sbaxdm.usc.edu/Publications/list.html).

Lee S., Dreze X. et Zufryden F. (2001), Modeling consumer learning and repeat visit behavior on the Internet, *Actes du Hawaii International Conference on System Sciences*, 34.

Lee S., Zufryden F. et Dreze X. (2000), A multivariate multinomial logit-Markov model and its application to consumer switching behavior on Internet, soumis au *Journal of Marketing Research*, http://sbaxdm.usc.edu/Publications/list.html.

Lendrevie J. (2000), Internet est-il doué pour la publicité? , *Revue Française du Marketing*, 177/178, 102-118.

Leontief W.W. (1986), *Input-output economics*, 2ième édition, Oxford University Press.

Levene M., Borges J. et Loizou G.(2001), Zipf's law for web users, *Knowledge and Information Systems: An International Journal*, 3, 1, 120-129.

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.

Lohse G.L., Bellman S. et Johnson E.J. (1999), Consumer buying behavior on the Internet: findings from panel data, *Journal of Interactive Marketing*, 14, 1, 15-29.

Lynch Jr J.G. et Ariely D. (2000), Wine online: Search costs and competition on price, quality and distribution, *Marketing Science*, 19, 1, 83-103.

Mandel N. et Johnson E.J. (1999), Constructing preferences online: Can web pages change what you want?, papier de recherche, MIT E-commerce forum, Wharton.

Moe W.W. et Fader P.S. (2000), Which visits lead to purchases? Dynamic conversion behavior at E-commerce sites, papier de recherche cité par Barwise P., Elberse A. et Hammond K.

Montgomery A.L. et Faloutsos C. (2000), Trends and patterns of www browsing behavior, papier de recherche, GSIA, 2000-E20, http://www.andrew.cmu.edu/user/alm3.

Downloaded from ram.sagepub.com at UNIVERSITE DU SUD TOULON VAR on May 31, 2013

Novak T.P. et Hoffman D.L. (1997), New metrics for new media: Toward the development of web measurement standards, *World Wide Web Journal,* 2, 1, 213-246.

Perkowitz M. et Etzion O. (2000), Towards adaptative web sites: Conceptual framework and case study, *Artificial Intelligence*, 118, 245-275.

Peterson R.A., Balasubramanian S. et Bronnenberg B.J. (1997), Exploring the implications of the Internet for consumer marketing, *Journal of the Academy of Marketing Science*, 25, 4, 329-346.

Pirolli P. et Pitkow J. (1999), Distributions of surfers' paths through the world wide web: Empirical characterizations, *World Wide Web*, 2, 1-2, 29-45.

Pirolli P., Pitkow J. et Rao R. (1996), Silk from a sow's ear: Extracting usable structures from the web, *Actes de la Conférence Human Factors in Computing Systems*, 118-125.

Pitkow J. (1995), In search of reliable usage data on www, *Actes de la 6ième International WWW Conference*, http://www.parc.xeros.com/istl/projects/uir/pubs/pdf/UIR-R-1997-04-Pitkow-www6-UsageData.pdf.

Pitkow J. et Kehoe C. (1995), Results from third world wide web user survey, http://www.w3.org/Conferences/WWW4/Papers/107/

Poulsen K. et Valette-Florence P. (1996), Improvements in means-end research: A heterogeneous latent Markov approach, *Marketing Science Conference*, Gainesville.

Scott J. (1991), *Social network analysis*, Newbury Park CA, Sage.

Shneiderman B. (1997), Designing information-abundant web sites: Issues and recommendations, *International Journal of Human Computer Studies*, 47, 1, 5-29.

Stout (1997), *Web site stats: Tracking hits and analyzing traffic*, Osborne, McGraw-Hill.

Tague J. et Nicholls P. (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.

Winne P.H., Gupta L. et Nesbit J.C. (1994), Exploring individual differences in studying strategies using graph theoretic statistics, *Alberta Journal of Educational Research*, 40, 2.

Xia S. et Yingzhong L. (1990), Multi-sectorial planing models, *Energy*, 3-4, 325-339.

Zipf G.K. (1949), *Human behavior and the principles of least effort: An introduction to human ecology*, Addison Wesley.

Downloaded from ram.sagepub.com at UNIVERSITE DU SUD TOULON VAR on May 31, 2013