NOUVELLES TECHNOLOGIES

Le webmining, pour bâtir une nouvelle relation client

***Les techniques de webmining** permettent de connaître le comportement de l'internaute afin de mieux répondre à ses attentes. Passage en revue des différentes méthodes, pistes de recherche et perspectives opérationnelles.*

PAUL DEMEY
ETIENNE MAROT
Ingénieurs d'études
ANTOINE FRACHOT
Responsable
Groupe de Recherche Opérationnelle Crédit lyonnais

Avec ses **8 150 000** internautes en France (1), ses 4 590 000 abonnés (2) et 11,1 % des foyers français équipés (3), internet est devenu un enjeu majeur de la politique CRM des entreprises. Grâce à la personnalisation des sites, à la vente on-line, au *cross-selling* et à tout ce que recèle l'e-CRM, internet est un canal de distribution supplémentaire de produits et services permettant une relation personnalisée avec le client et un suivi précis de son activité. Ce suivi, reposant le plus souvent sur une analyse de trafic et sur du tracking, aboutit à des règles opérationnelles destinées à offrir à l'internaute une configuration optimale en termes d'ergonomie, et à l'entreprise une rentabilité maximale de son site web.

Dans le domaine bancaire, les sites de consultation des soldes, de simulation d'épargne, de gestion de portefeuille et de gestion des comptes présentent un intérêt particulier : étant donné qu'ils requièrent une identification du client, ils permettent de recouper les données issues du parcours de l'internaute dans le site avec les données commerciales de la banque, ce qui ouvre de nouvelles perspectives en termes de gestion de la relation client, de prospection, et globalement de connaissance de la clientèle.

Les méthodes informatiques et statistiques permettant de suivre précisément l'activité d'un site web ou d'en améliorer l'ergonomie et la rentabilité sont regroupées sous le terme général de *webmining*. Le *webmining* commence par une simple description des données disponibles (nombre de pages visitées, nombre de connexions dans une période définie…) et peut aboutir à une modélisation des relations entre les différentes pages du site.

LA STRUCTURE DES DONNÉES

Les types de données disponibles pour chaque site web sont de deux sortes :

- les *cookies* sont des documents textes stockés sur l'ordinateur de l'internaute qui contiennent un identifiant numérique permettant au serveur de reconnaître l'internaute lors de ses connexions et ainsi de charger son «profil». C'est le premier élément de relation personnalisée sur le web, utilisable en temps réel par le serveur, mais inexploitable par la suite ;
- la *log* est un fichier texte stocké sur le serveur, enrichi d'une ligne à chaque fois qu'un élément (page, image, script, fichier à télécharger…) du site est requis par un utilisateur. C'est sur ce type de fichier que toutes les analyses statistiques sont fondées.

Il existe différents types de *log* (CLF, XLF, applicative) qui permettent de faire des analyses différenciées en fonction du niveau d'information dont on dispose sur les internautes. Nous présentons ici les analyses possibles avec un niveau d'information disponible croissant : étude des parcours dans le site, estimation des transitions et règles d'association, *profiling*, caractérisation d'une typologie des internautes, et finalement construction de scores d'appétence.

RÈGLES D'ASSOCIATION…

Les logiciels classiques d'analyse de logs web (Webtrends, Weborama, Sawmill, LiveStats…) fournissent pour une période donnée un grand nombre d'indicateurs de trafic, parmi lesquels le nombre de pages consultées, le nombre de connexions, les pages d'entrée, de sortie, les sites de provenance, le

> **« Les règles d'association permettent d'améliorer l'ergonomie d'un site web. »**

nombre de pages vues, les fichiers le plus souvent téléchargés, les parcours les plus empruntés… Ces indicateurs ne font que décrire des données disponibles.

Les règles d'associations permettent de prendre en compte une dimension supplémentaire dans cette analyse : la corrélation entre les pages visitées. Ce sont des règles du type «60 % des internautes qui visitent la page *p* visitent également la page *q*».

Pour un site donné, il existe presque autant de règles qu'il y a de combinaisons de pages. Ce grand nombre de possibilités à explorer engendre des problèmes algorithmiques dont la résolution a fait l'objet de nombreuses études.

L'intérêt opérationnel des règles d'association est double :

- repérer les liens logiques entre les visites de pages, sans notion de chronologie, et donc optimiser la répartition du contenu (par exemple en termes de liens vers d'autres sites web ou d'espace publicitaire) dans une perspective d'augmenter la rentabilité du site ;
- revoir l'ergonomie du site en termes de circulation pour accroître le confort de l'internaute.

« Il est apparu qu'on pouvait optimiser un site web pour accroître le PNB qu'il génère. »

… ET PROBABILITÉS DE TRANSITION

Si, à la notion d'«association», on veut ajouter une notion temporelle, il est nécessaire de considérer les probabilités de transition entre les pages. On obtient alors des règles du type : «35 % des internautes qui visitent la page *p* visitent, immédiatement après, la page *q* puis la page *r*».

Pour calculer ces règles, on utilisera une matrice de transition qui regroupe toutes les probabilités de passage d'une page à une autre. Cette analyse mène également à la découverte de chemins fréquents dans le parcours des internautes sur le site.

Ces méthodes (analyse du trafic, règles d'association, probabilités de transition) permettent de décrire le comportement global des internautes ; intéressons-nous maintenant à son comportement individuel.

LE PROFILING, POUR CONNAÎTRE L'INTERNAUTE

Le *profiling* consiste à décrire avec le plus de précision possible le comportement ou le potentiel de consommation d'un individu et donc de mieux répondre aux attentes d'un internaute, en lui proposant une ergonomie personnalisée, des liens adaptés à ses centres d'intérêt, des produits et services qu'il est susceptible d'acheter, des publicités auxquelles il est sensible… Profiler un site est donc un excellent moyen de le rendre interactif à l'insu de l'internaute.

Sur internet, on entend le plus souvent par *profiling* la construction et l'enrichissement d'un vecteur de préférences pour chaque internaute. Ce vecteur de préférences est construit à l'aide d'une classification des rubriques d'un site et du nombre de «hits» enregistrés sur les rubriques pour un utilisateur ; une simple mise à jour du vecteur de l'utilisateur permet de redresser son profil et de lui proposer, à sa prochaine connexion, une nouvelle interface personnalisée plus en harmonie avec son nouveau profil.

Pour les sites sans identification de l'utilisateur, cette technologie repose sur l'existence des *cookies*, qui permettent au serveur de reconnaître dès le début de la connexion l'individu auquel il a affaire. Néanmoins, un *cookie* n'identifie pas un individu, mais un ordinateur, et il peut être supprimé par l'utilisateur. Le profil ainsi construit et enrichi est celui d'un foyer pour les ordinateurs domestiques, et il peut être réinitialisé à tout moment.

Le fonctionnement du *profiling* est le suivant : à sa première connexion sur un site, l'internaute reçoit sur son ordinateur un *cookie* du nom du site lui donnant un identifiant numérique. En même temps, son profil est créé, avec une composante de 100 % en finance si la page visitée est du thème «finance». À la page ou à la connexion suivante, le profil est actualisé avec les nouveaux thèmes visités, en proportion du nombre de pages vues : pour 6 pages de finance et 4 pages de simulation d'épargne visitées sur 10 pages, le profil sera à 60 % «finance» et à 40 % «épargne».

Symétriquement, lorsqu'un utilisateur se connecte à un site personnalisé, l'identifiant contenu dans le *cookie* est communiqué au serveur, qui peut dès lors proposer à l'internaute un contenu et des liens personnalisés. Cette personnalisation présente de nombreux avantages :

- elle est éthique : elle n'utilise aucune donnée socio-démographique et est conforme aux exigences de la CNIL ;
- elle est transparente pour l'internaute, qui n'a aucune information à fournir : la seule source de données est son comportement sur le web ;
- elle est rapide : un profil étant un simple vecteur, il est aisé de le mettre à jour et de personnaliser les pages en temps réel ;
- elle est très évolutive : la moindre page visitée a une influence sur le vecteur de composantes ;
- elle reflète le comportement récent de l'internaute : de manière naturelle, l'importance relative des pages rarement visitées est rapidement diminuée.

En revanche, elle trouve ses limites dans les limites techniques des *cookies* : ils peuvent être refusés (4) ou détruits par l'utilisateur. Elle est en outre difficile à mettre en œuvre.

Créer une typologie des internautes

Créer une typologie de ses internautes demande la connaissance d'indicateurs fiables de leur activité. Ceci ne peut s'appliquer qu'aux sites requérant une identification de l'internaute compatible avec ses informations bancaires. Il est alors aisé de construire des classes d'internautes en fonction des critères qui intéressent la banque, selon l'intensité de visite du site, le nombre de rubriques visitées, d'opérations bancaires on-line, le montant des opérations boursières on-line.

La constitution de ces classes fait appel aux techniques de classification classiques ; la plus-value vient de la caractérisation des classes ainsi construites par des variables supplémentaires (âge, PCS, situation familiale, équipement bancaire…). Le profil des internautes prend dans ce cas une autre dimension : c'est aussi un profil de client de la banque, et il peut à ce titre être utilisé dans des opérations marketing de promotion d'un produit, servir pour améliorer des modèles de score, ou compléter des analyses de la structure de la clientèle.

On peut par exemple déduire de cette analyse des pistes intéressantes pour la construction de scores d'appétence ou d'affinité destinés soit à compléter l'équipement en produits et services des internautes, soit à vendre des produits internet aux clients dont le profil est similaire à celui des abonnés et qui ne sont pas encore équipés.

Des méthodes perfectibles

Ces différentes méthodes, les plus utilisées en *webmining*, méritent des améliorations : notre expérience de l'ensemble de ces méthodes a mis en évidence certaines failles et pistes d'exploration, parmi lesquelles la détermination statistique d'un seuil de significativité pour les règles d'association (la structure du site ne doit-elle pas être prise en compte ?), la mesure de l'information apportée dans le site par une page en termes de navigation (par rapport à un parcours aléatoire), le classement des internautes selon un critère d'«originalité» de ses parcours (distance de son parcours moyen au parcours moyen de tous les internautes), la mesure de la rentabilité d'un site bancaire en termes de PNB (avec une mesure du gain en PNB généré par chacune des pages)… Autant de pistes qui font du *webmining* un domaine de recherche incontournable. ■

(1) Médiamétrie, décembre 2000. Un internaute est une personne s'étant connectée à internet au moins une fois dans les douze derniers mois.
(2) AFA, septembre 2000.
(3) Médiamétrie/ISL, 3e trimestre 2000. À la même époque, 33,2 % des internautes déclaraient se connecter une fois par jour ou presque.
(4) Dans le droit français, l'installation d'un *cookie* à l'insu de l'internaute peut tomber sous le joug de l'article 323 du Code pénal. En théorie, il convient de demander à l'utilisateur la permission d'installer un *cookie* sur son ordinateur. Cependant, même si les navigateurs ont une option qui permet de refuser, confirmer ou automatiser l'écriture des *cookies*, la plupart les acceptent tous par défaut.