Chapitre 4

Décrire et visualiser la dynamique des parcours

L'ancrage temporel et séquentiel des parcours sur le Web nous interdit de nous limiter à l'analyse des contenus visités : de la même manière qu'un texte n'est pas un sac de mots, une session de navigation n'est pas une collection de pages, mais un cours d'action où la visualisation de chaque page prend sens dans la dynamique générale du parcours. La sémantique des parcours se doit de rendre compte de cette dynamique : pour cela, nous proposons à la fois de recourir à des outils de fouille manuelle pour travailler au plus près de l'objet, ainsi que des éléments de représentation statistiques de la topologie des parcours. Pour cela, nous opposons aux travaux menés jusqu'alors sur la navigation une approche descriptive se situant dans le cadre des sciences humaines, dans le cadre d'une théorie de l'action. À ce titre, la session doit être replongée dans le contexte individuel de chaque internaute : ses pratiques du Web, ses usages d'Internet en général, pour autant que les déterminations globales de l'activité de navigation influencent directement les éléments locaux – forme, contenu.

4.1 Outils de fouille des données

L'analyse de données de trafic volumineuses ne doit pas sacrifier au « tout statistique » : il est indispensable de pouvoir mener un examen manuel des parcours pour en appréhender la logique et la complexité. Pour répondre à ce besoin, nous avons développé deux outils de manipulation des parcours qui trouvent naturellement leur place dans l'outillage d'analyse de la navigation.

4.1.1 Rejouer les parcours

Nous avons, face aux données de trafic, rapidement éprouvé le besoin de pouvoir « refaire » des parcours, c'est-à-dire de revoir dans l'ordre de la visite les différentes URL composant une session. Nous avons pour cela développé un outil baptisé *RePlay*

qui permet, pour une session donnée, de visiter l'ensemble de ses pages dans l'ordre et la temporalité d'origine.

L'application consiste à fournir une interface HTML pour reproduire la session dans la fenêtre d'un navigateur (voir Figure 4.1). Utilisant le mécanisme des *frames*, elle propose, dans une partie gauche de l'écran, la liste ordonnée des noms de sites visités dans une session ; chacun de ces éléments contient un lien hypertexte vers l'adresse de la page vue par le panéliste, qui s'affiche dans le *frame* de droite.

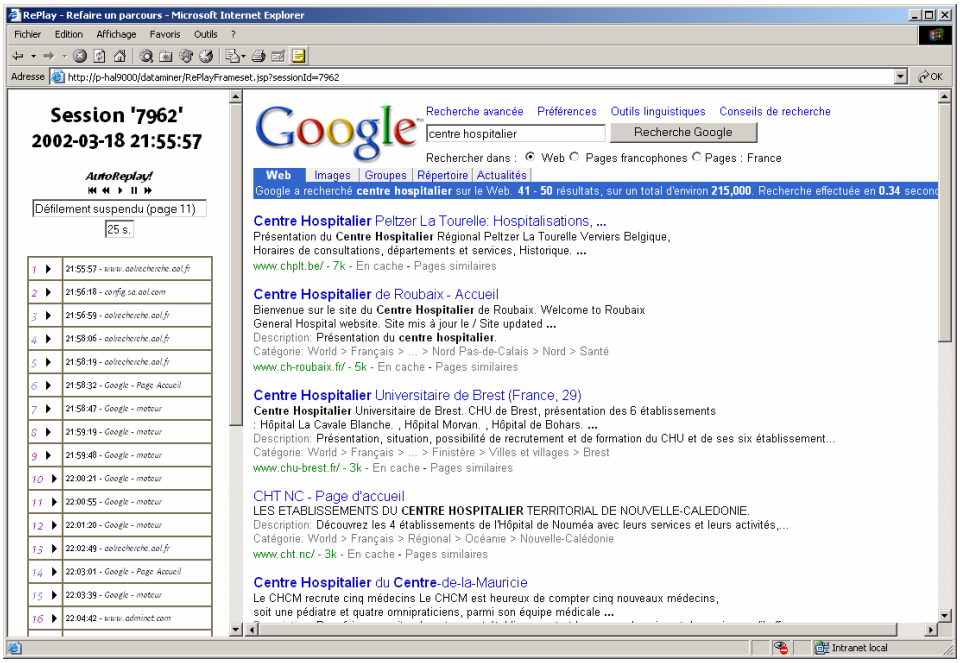

Figure 4.1. Interface de RePlay – vue générale

Deux modes d'utilisation sont proposés : dans le premier, l'utilisateur de *RePlay* clique sur chaque lien pour ouvrir la page visitée par l'internaute. Cet accès permet d'aller rapidement aux passages jugés intéressants et de vérifier rapidement certains contenus, ou de passer du temps sur d'autres. Le deuxième mode reproduit automatiquement les requêtes une à une en respectant les délais observés entre deux requêtes par l'internaute. L'utilisateur de *RePlay* regarde ainsi « défiler » la session sous ses yeux, et bénéficie de l'effet de temps passé sur chaque page.

Ces deux modes ne sont pas antagonistes : il est possible de lancer le défilement automatique à partir de n'importe quelle page de la session, et de l'interrompre à tout moment. Une console de commande *AutoRePlay* (voir Figure 4.2 ci-dessous) permet de contrôler le défilement, de le suspendre ou de le relancer à tout moment, et de forcer le passage à la page suivante, sur le mode de la lecture des pistes d'un CD audio. En outre, cette console affiche l'URL visitée, et le temps total que l'internaute a passé dessus ; certains temps de visualisation pouvant être très longs (en théorie, jusqu'à trente minutes étant donné le mode de calcul des sessions), la durée effective entre deux pages a été positionnée à 30 secondes maximum en mode défilement.

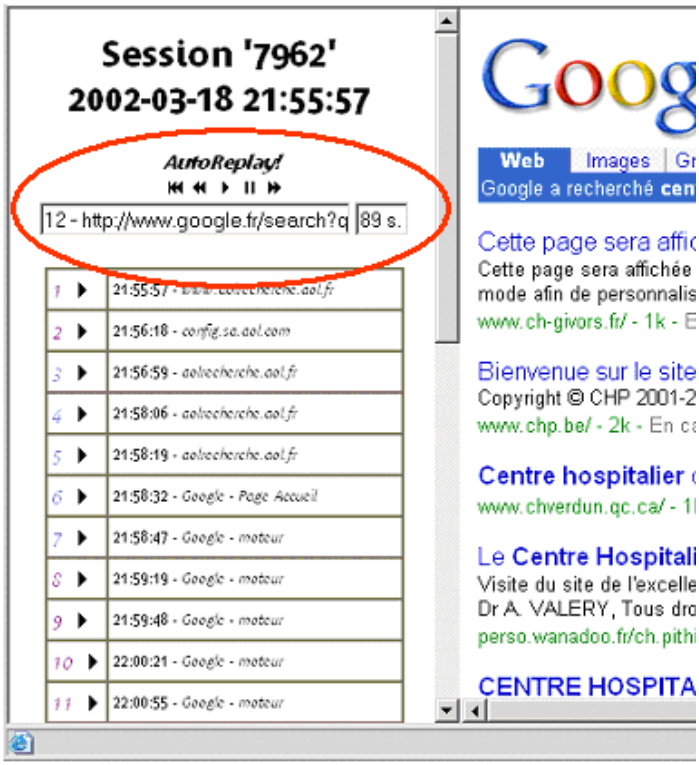

Figure 4.2. Interface de RePlay – détail

Bien évidemment, cette interface ne fait pas abstraction du problème du renouvellement ou de la disparition des pages Web, dans la mesure où *RePlay* effectue une requête en différé vers des URL accédées plusieurs mois auparavant ; en outre, on se heurte également au problème des accès restreints et authentifiés. Bref, nous rencontrons ici les mêmes aléas que lors de l'aspiration de pages (voir chapitre 3.2). En outre, le problème des *frames* n'est pas résolu par ce mode de représentation : comme nous l'avons vu au Chapitre 2, l'unité ergonomique perçue par l'utilisateur peut être le fruit de plusieurs requêtes, notamment lorsque la page affichée contient des pages imbriquées par le mécanisme des *frames*. Dans ce cas, *RePlay* effectue une requête par élément, et affiche le contenu de chaque *frame* séparément, ce qui nuit à la lisibilité globale du parcours.

Cet outil a été développé sous la forme de servlets Java qui interrogent directement la base de données de trafic, pour être intégré à la plateforme de fouille de données de trafic SensNet, et reconstruit dynamiquement une interface à chaque session demandée. Cette application nécessitant d'être connecté à cette base, une version *stand-alone* a été développée, qui permet, sous la forme de deux fichiers HTML, de transporter les résultats de l'application et de refaire les parcours depuis n'importe quelle machine disposant d'un accès Internet.

> ***Synthèse.*** *Le module* RePlay *permet de « rejouer » une session en revoyant en différé le contenu d'un parcours dans l'ordre et la temporalité de la visite par l'internaute. Si le module se heurte aux problèmes de l'évolution*

des pages et de l'accès restreint, il est néanmoins un puissant outil pour formuler et vérifier des hypothèses sur les parcours, et mettre à jour les logiques à l'œuvre dans les sessions.

4.1.2 Représentation graphique

L'intérêt d'une représentation synthétique des sessions sous forme graphique est double : d'une part, elle permet de formuler et de vérifier des hypothèses sur les parcours et les liens entre forme et contenu. D'autre part, la visualisation graphique d'une navigation a une valeur didactique qui n'est pas à négliger dans la présentation des travaux sur les parcours.

Pour répondre à ce besoin, nous avons développé deux outils capables de représenter les sessions sous la forme d'un graphe : dans les deux cas, les nœuds du graphe sont les pages ou les sites visités, et les arcs orientés représentent le passage d'une page ou d'un site à l'autre. Afin de rendre les graphes ainsi obtenus plus lisibles, nous avons représenté différemment les URL ou les sites taggués comme services dans les portails généralistes : ce n'est alors plus l'adresse, mais le nom du portail et le service, dans le cas d'un graphe d'URL, qui sont affichés. Quatre types de graphes sont produits, en fonction de l'échelle d'analyse à laquelle on se place :

- graphe de pages : les nœuds représentent les URL visitées ;
- graphe de pages/services : les nœuds représentent les URL visitées, mais les URL correspondant à un service sur un portail identifié par *CatService* sont regroupées et représentées comme telles, et sont colorées ; en outre, s'il s'agit d'un moteur de recherche, le contenu de la requête est affiché, et les différentes pages de résultat sont regroupées ;
- graphe de sites : les nœuds représentent les sites visités, et chaque nœud agrège l'ensemble des URL visitées sur un même site ;
- graphe de sites/services : comme dans le cas précédent, les nœuds représentent les sites, mais si ce site est un portail identifié par *CatService* et qu'un service est spécifié, on distingue dans le graphe les différents services utilisés sur le portail, qui donnent lieu à autant de nœuds différents.

Ne pouvant ni ne souhaitant développer un algorithme de mise en forme des graphes – champ de recherche qui dépasse de très loin notre étude – nous avons fait appel à deux solutions existantes pour la mise en forme.

Solution Graphlet

Le premier outil employé pour la mise en forme de graphes de sessions est le logiciel Graphlet[1]. Notre outil extrait des données de trafic, en fonction de la granularité souhaitée, la liste ordonnée des URL ou des sites d'une session et la date de chaque requête ; à partir de cette liste, il prépare un fichier au format supporté par Graphlet, où sont spécifiés les nœuds, les arcs, leurs labels, les couleurs des nœuds,

[1] Logiciel développé par l'Université de Passau (Allemagne) ; voir `http://www.infosun.fmi.uni-passau.de/Graphlet/`.

mais aucune coordonnée ; il faut ensuite ouvrir le fichier dans Graphlet et appliquer à notre graphe un des algorithmes de mise en forme proposés.

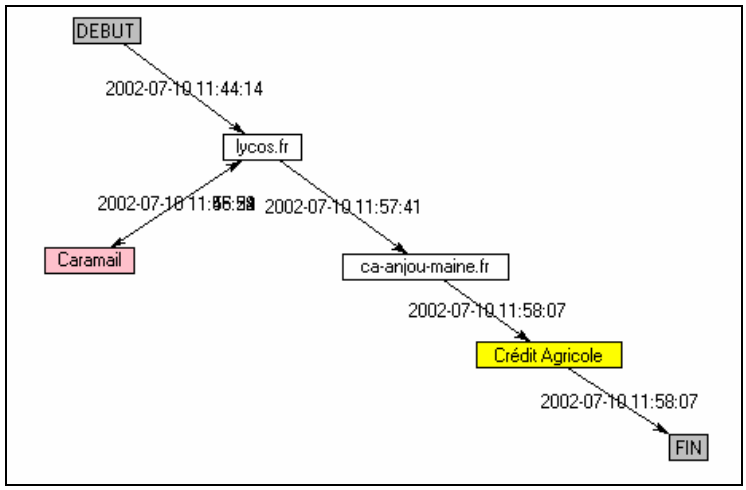

Figure 4.3. Exemple de graphe de session à l'échelle du site (SN2002, session 127666)

La Figure 4.3 présente un exemple de sortie obtenue après ce traitement, au niveau de granularité du site. La représentation permet d'apprécier la grande différence en termes de linéarité entre l'analyse à l'échelle de la page et à celle du site : la Figure 4.4 représente la même session, au niveau de la page cette fois. On constate que, si une linéarité globale se dessine, quatre « déviations » sont opérées, contre une seule au niveau des sites.

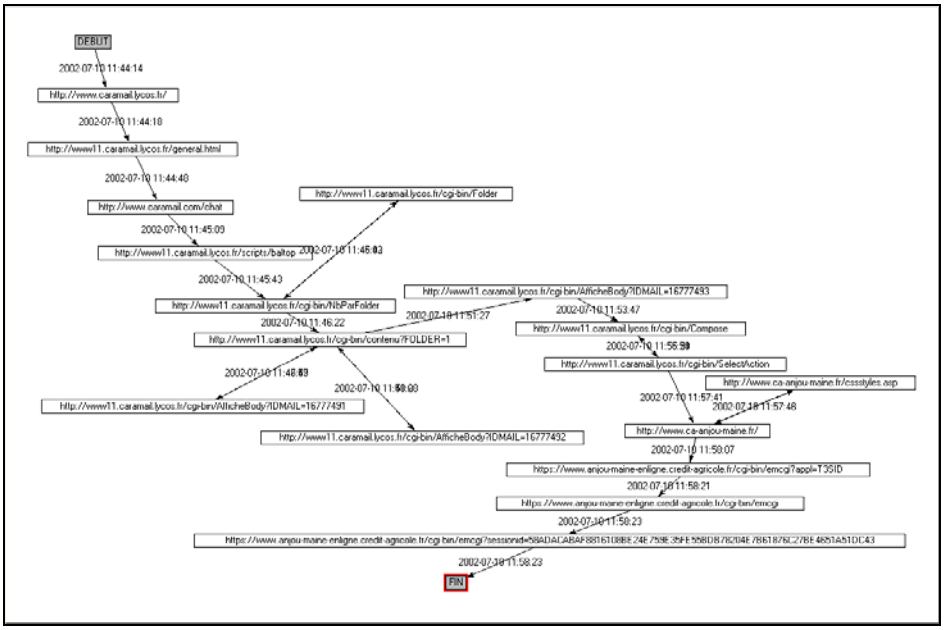

Figure 4.4. Graphe de session à l'échelle de la page (SN2002, session 127666)

Lorsque le nombre de nœuds distincts est supérieur à une dizaine, la lisibilité des graphes se trouve quelque peu altérée, mais les représentations restent cependant exploitables. La Figure 4.5 ci-dessous montre un exemple de graphe au niveau site particulièrement dense, avec beaucoup de sites visités et de retours en arrière, certains sites agissant comme de véritables pivots dans la navigation. La représentation qui en découle est touffue, pas toujours lisible, mais on peut néanmoins dégager une thématique générale à la session : dans cet exemple de session datée du 21 avril 2002, on voit clairement que, outre les portails généralistes et les moteurs, les sites visités sont liés à la politique et aux élections présidentielles. L'effet de balayage des ressources est très fort, l'ensemble des tendances politiques étant représentées.

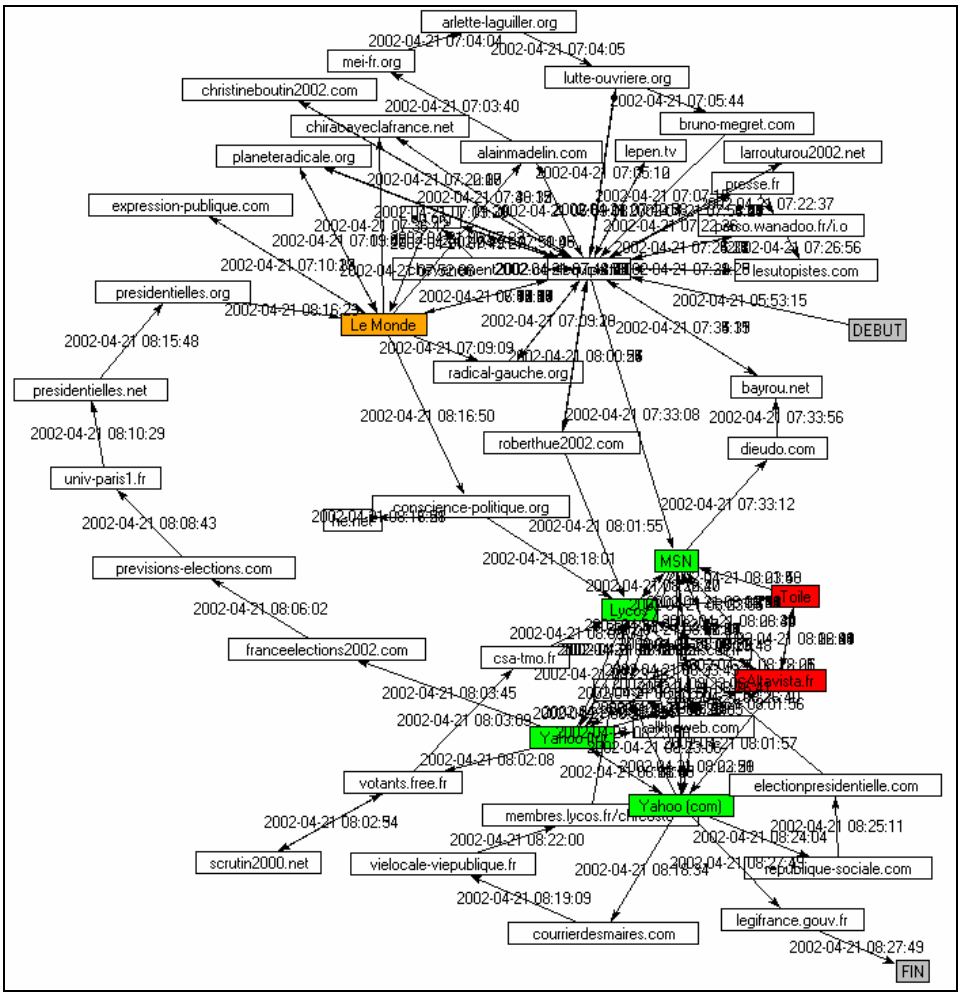

Figure 4.5. Graphe de session à l'échelle du site (SN2002, session 2461)

L'avantage de Graphlet est de proposer plusieurs algorithmes avancés de mise en forme ; le graphe produit peut être retouché finement, au niveau des labels, de la taille, du format des nœuds, etc. et le résultat peut être enregistré au format GML

(format Graphlet). Par contre, la production des graphes n'est pas automatisée : Graphlet ne peut pas être utilisé en ligne de commande, et il faut éditer chaque graphe non mis en forme dans l'interface graphique, appliquer un algorithme et le sauvegarder ensuite, ce qui est fastidieux lorsque l'on veut visionner un nombre important de graphes.

Solution Java

La deuxième solution à laquelle nous avons recouru pallie ce problème d'automatisation : le graphe est ici présenté dans une *applet* Java (voir Figure 4.6). Les paramètres relatifs aux différents nœuds et arcs du graphe sont passés à l'*applet* dans le fichier HTML qui l'appelle, ce qui permet ici aussi de faire une version *stand-alone* de l'outil, contenant les classes Java nécessaires et le fichier HTML relatif à la session représentée.

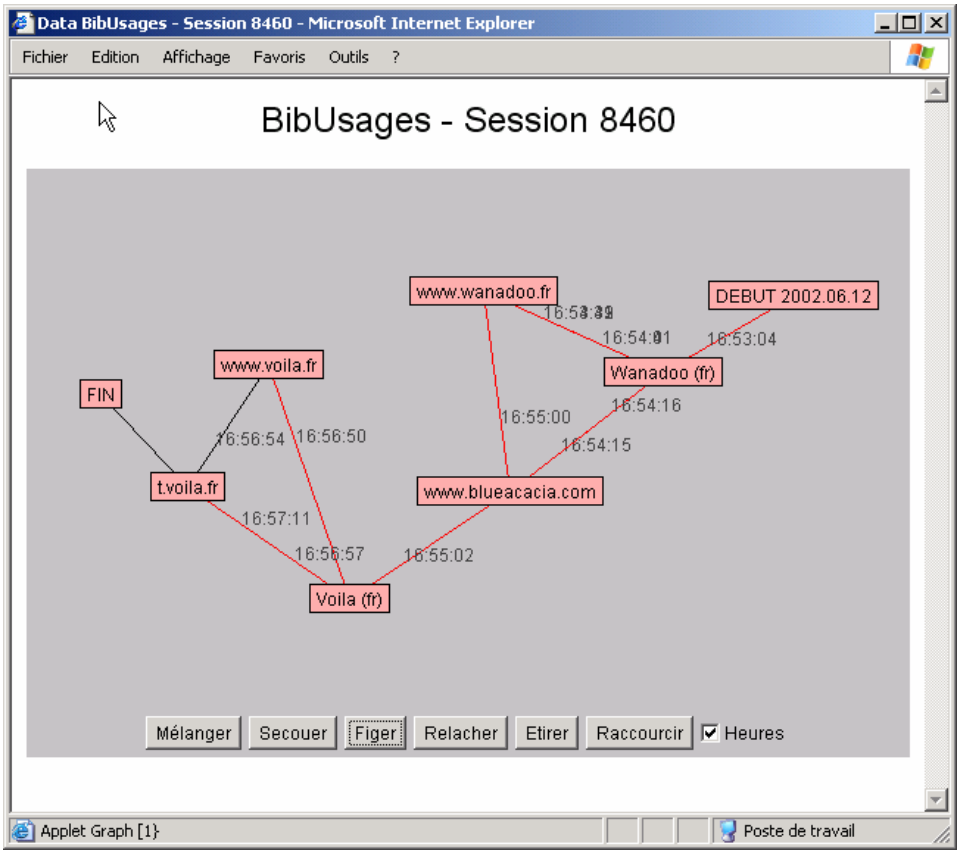

Figure 4.6. Graphe de session à l'échelle du site - Applet Java

L'inconvénient de cette solution est qu'elle ne propose pas un mécanisme de mise en forme élaboré : l'algorithme utilisé[1], assez rudimentaire, se base sur les distances entre nœuds, et tente d'étaler le graphe au mieux. Pour autant, le résultat est satisfaisant pour des graphes de faible dimension, d'autant plus que le graphe est facilement manipulable : l'utilisateur voit la mise en forme se faire dans l'*applet*, et peut intervenir pour déplacer un nœud, le figer, allonger ou raccourcir la distance entre les nœuds, et figer l'ensemble ou laisser se poursuivre la mise en forme. Cette souplesse compense en partie la faiblesse du procédé d'élaboration du graphe.

L'autre intérêt particulier de cette solution en Java est son interfaçage complet avec les données de trafic et les outils de fouille développés dans le projet SensNet. Là où il faut avec Graphlet générer un fichier GML, l'ouvrir puis le mettre en forme pour chaque session, ce qui est fastidieux, le graphe est produit dynamiquement dans une interface Web et sa production est totalement transparente pour l'utilisateur, qui fait l'économie des manipulations techniques.

Dans les deux cas, nous voyons l'agrément de cette représentation des parcours en même temps que ses limites : les graphes sont clairs lorsque les parcours sont assez courts et linéaires, mais rapidement illisibles lorsque les sessions s'allongent. En outre, nous rencontrons ici des problèmes inhérents aux données sur lesquelles nous travaillons, à savoir la différence entre ce que perçoit l'utilisateur (l'unité visuelle de la page) et les requêtes multiples qui peuvent en être à l'origine. Ainsi, comme nous l'avons vu pour *RePlay*, l'utilisation de *frames* par les concepteurs de sites génèrera au moins trois requêtes, et trois nœuds sur le graphe, alors qu'il s'agit, du point de vue de l'utilisateur, d'une seule page. Cela étant, l'outil s'est montré très utile et a globalement répondu à nos attentes.

À l'usage, cet ensemble d'applications de fouille minutieuse des données s'est avéré précieux, en ce qu'il rend plus palpables les logiques de navigation en les replaçant dans un pseudo-contexte d'utilisation. Il a également, et c'est son principal intérêt, permis de formuler certaines hypothèses, parmi lesquelles :

- le comportement des utilisateurs varie grandement en fonction des services accédés, en particulier en termes de longueur de session (temps et nombre d'URL).
- on tend à observer une forme d'opposition entre comportement « prédateur » (courte session, le panéliste sait où il va ou ce qu'il cherche avec précision) et « fureteur » (session plus longue, plus diversifiée, où le suivi des liens entre pages semble tenir d'un certain opportunisme).
- la session n'apparaît pas comme une unité cohérente du point de vue du contenu, et l'on observe dans un certain nombre de sessions des formes de « coq à l'âne » assez radicaux.
- la vision très « mono-tâche » de la navigation Web, que l'on retrouve dans un certain nombre d'études, est mise à mal par l'observation d'entrelacements au sein d'une même session de deux, voire plus, cours

[1] Cet algorithme est emprunté aux exemples d'*applets* proposés par Sun ; voir `http://java.sun.com/applets/jdk/1.1/demo/GraphLayout/`.

d'action distincts, qui peuvent correspondre à l'utilisation de plusieurs fenêtre du navigateur utilisées simultanément.

Toutes ces hypothèses sont loin d'être exhaustives, et méritent bien évidemment d'être vérifiées. Elles traduisent surtout la démarche hypothético-déductive qui est la nôtre, et qui implique la possibilité d'aller et venir entre des représentations synthétiques de masse et un examen minutieux des données de navigation. C'est dans ce cadre que l'élaboration d'indicateurs statistiques rendant compte ce que nous percevons lors de la fouille minutieuse des données s'avère précieuse : face aux données volumineuses de trafic dont nous disposons, l'examen manuel systématique est impossible, et nous devons nous doter d'outils synthétiques de représentation des contenus et des formes de parcours pour la vérification massive des observations manuelles.

> ***Synthèse.*** *La représentation des parcours sous forme de graphes permet, comme* RePlay, *d'émettre et de vérifier des hypothèses sur les parcours, et alimente une approche hypothético-déductive. Elle offre également une vue synthétique des sessions à différentes échelles (page, site, service) qui permet d'appréhender leur complexité : détours, retours arrière, pages-pivots, etc. Une telle vue est particulièrement utile pour l'analyse de la topologie des parcours.*

4.2 Analyser la séquentialité

Si les outils de fouille permettent d'approcher au plus près les données de navigation et de formuler des hypothèses de travail, l'analyse de données de trafic volumineuses nécessite la construction de représentations et de descriptions synthétiques des parcours. Ces descriptions permettent de vérifier les hypothèses sur les comportements de navigation à l'aide de traitements statistiques et formels sur les parcours, et ouvrent la voie vers l'élaboration de profils-types de sessions et d'outils de classification automatique des parcours.

4.2.1 Parcours Web : travaux existants

L'existence de nombreux systèmes hypermédia avant le Web a donné lieu à nombre de travaux sur la navigation, en particulier dans le champ des sciences cognitives. Les recherches portent alors principalement sur la modélisation de l'utilisateur à travers l'étude de ses parcours dans un système hypermédia donné ; les applications sont alors tournées vers les recommandations de conception et surtout la mise en place d'hypermédias adaptatifs (*adaptive hypermedia*). On trouvera dans [Brusilovsky 1996] un panorama très complet des problématiques, des méthodes et des applications relatives aux hypermédias adaptatifs avant l'émergence du Web, complété en 2001 dans [Brusilovsky 2001] pour les études centrées sur le Web. Les champs d'applications principaux sont centrés autour des sciences de l'éducation (application à des encyclopédies, des méthodes d'apprentissage multimédia) et à la recherche d'information, proche de l'ingénierie documentaire.

Dans ce cadre de recherche, les éléments qui nous intéressent ont trait au travail descriptif nécessaire à la modélisation des comportements d'utilisateurs : d'une part les méthodologies et indicateurs construits pour rendre compte des navigations dans les systèmes hypermédias, et d'autre part les différents profils-types qui ont pu être extraits des observations. Les modèles construits proprement dits ne retiennent pas notre attention : nous ne visons pas à modéliser les comportements pour faire de la prédiction, mais à disposer d'outils permettant de décrire et d'analyser les comportements observés dans le cadre de pratiques en situations réelles.

Browsing patterns et modélisation de l'utilisateur

Ceci étant posé, que retiendrons-nous des travaux orientés *user modeling* ? En premier lieu, au niveau minimal des actions de navigation, Canter, Rivers et Storr proposent dans [Canter *et al.* 1985] d'identifier quatre formes de bases dans la navigation (voir Figure 4.7 ci-dessous) :

- *pathiness* (chemin) : un chemin est un parcours qui ne passe pas deux fois par le même nœud ;
- *ringiness* (anneau) : un anneau est un parcours qui retourne à son point de départ ;
- *loopiness* (boucle) : une boucle est un anneau qui ne contient pas de sous-anneau ;
- *spikiness* (pointe) : une pointe est un anneau qui retourne à l'origine en repassant par les nœuds intermédiaires.

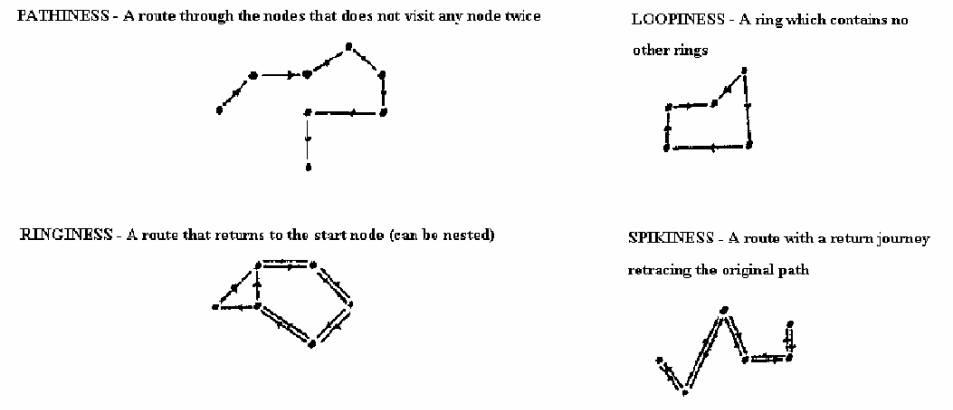

Figure 4.7. Formes de bases de navigation présentées dans [Canter et al. 1985]

Ces quatre formes sont dénombrées dans les sessions de navigation, à quoi s'ajoutent deux ratio : 1) le ratio du nombre d'éléments visités rapporté au nombre total d'éléments, et 2) le ratio du nombre d'éléments visités rapporté au nombre total de visites. Les auteurs obtiennent ainsi six indicateurs pour représenter de manière formelle la navigation, et les analysent pour dégager des modèles de navigation-type. Notons ici que la reconnaissances de ces quatre formes est loin d'être triviale, en particulier l'extraction des *boucles* et des *anneaux* dont les imbrications peuvent être complexes ; [Mullier 2000] propose ainsi une méthode de reconnaissance automatique de ces formes sur la base de réseaux de neurones.

Ce type de recherches sur les hypermédias a trouvé un prolongement naturel et productif sur l'hypertexte particulier que constitue le Web : on trouve une littérature relativement abondante traitant de l'analyse des parcours d'utilisateurs d'un point de vue *site-centric*, sur la base de l'analyse des *logs* des serveurs Web. S'appuyant sur les techniques mentionnées ci-dessus ou mobilisant des analyses statistiques sur la base de chaînes de Markov, d'analyse de séries temporelles ou d'outils de *data mining*, les travaux cherchent à découvrir des motifs récurrents de navigation (*browsing patterns*) sur un site donné. Les applications sont essentiellement orientées vers la conception (amélioration de l'architecture et de l'ergonomie d'un site), l'analyse de la fréquentation (rubriques les plus visitées, segmentation des sessions par type de visite), l'optimisation des serveurs et l'élaboration de contenus adaptatifs (prédiction, proposition de liens et contenus personnalisés à la session). Un tel engouement s'explique par les enjeux économiques sous-jacents à ces recherches : les sites à vocation commerciale souhaitent disposer de données les plus précises possibles sur leur fréquentation, afin de savoir quelles pages sont les plus visitées, comment les utilisateurs y arrivent, et comment les faire « rester » plus longtemps sur le site.

Toute une série de travaux a suivi cette voie en conservant les paradigmes issus des études sur les hypermédias ; ces recherches se situent dans le champ des sciences cognitives et s'orientent vers la modélisation de l'utilisateur en situation de navigation sur le Web. Nous renvoyons à la lecture de [Modjeska 1997] pour un panorama des travaux effectués dans ce domaine, certes un peu daté, mais qui rend bien compte des problématiques soulevées par cette approche, qui fait la part belle aux perceptions, aux « structures cognitives » et aux « modèles mentaux » de l'utilisateur.

Dans la plupart des cas, il s'agit d'études centrées-utilisateur sur des panels restreints, parfois tournées vers l'« usabilité » d'un site en particulier, le problème de la « désorientation » des utilisateurs[1] mais le plus souvent orientées vers la recherche d'information. Ce paradigme, directement hérité de l'ingénierie documentaire, domine encore la recherche sur la navigation Web, où les contenus sont assimilés à des documents contenant des « molécules informationnelles », avec en arrière-plan une vision orientée « exécution de tâche » et résolution de problème (problématique héritée de l'Intelligence Artificielle).

Nous ne passerons pas en revue l'ensemble de ces études, mais proposons d'en décrire une qui nous paraît particulièrement représentative des problèmes qui sont posés dans ce cadre. Il s'agit du travail de D. Mullier, D. Hobbs et D. Moore ([Mullier *et al.* 2002]) : sur la base des indicateurs de [Canter *et al.* 1985], les auteurs appliquent une méthode basée sur des réseaux de neurones pour reconnaître ces motifs dans des données de navigation dans un hypermédia ([Mullier 2000]), « et les interpréter (lorsque c'est possible) ». Ce système est appliqué à l'analyse de la navigation en situation de recherche d'information : on demande à onze étudiants d'effectuer une recherche sur un thème donné (l'astronomie) pour répondre à une question précise (quel est le plus gros satellite du système solaire), puis de naviguer à leur guise dans ce domaine dont ils n'ont pas tous la même connaissance. Les auteurs

---

[1] Problème du type « lost in hyperspace », abordé sous l'angle des modèles mentaux côté site et côté utilisateur ; voir en particulier [Xu *et al.* 2001] et [Danielson 2003].

ont notamment observé des motifs topologiques différents en fonction de l'expertise des volontaires, les experts présentant des navigations avec plus de boucles que les autres.

Les conclusions sont alléchantes, rejoignent nos problématiques et nous intéressent directement. On peut toutefois reprocher à cette étude de ne pas donner plus d'informations sur les contenus visités : quels sites sont vus, leur nombre, les utilisateurs connaissaient-ils les sites qu'ils ont visités (ce qui influence directement leur navigation dans ces sites), etc. Faute d'apporter des précisions sur ces éléments, l'étude se révèle peu convaincante.

Mais plus encore, c'est la question de l'interprétation des motifs observés qui pose problème à nos yeux : les travaux effectués dans ce champ amènent la plupart des auteurs à définir des taxinomies de « stratégies de navigation ». Il est ainsi souvent fait référence aux quatre classes définies dans [Canter *et al.* 1985] :

- *scanning* (feuilletage) : l'utilisateur passe en revue un nombre important de pages sur un thème donné sans passer beaucoup de temps sur chaque page, de manière superficielle ;
- *browsing* (navigation) : l'utilisateur suit un chemin jusqu'à parvenir à son but ;
- *searching* (recherche) : l'utilisateur cherche un document ou une information en particulier ;
- *exploring* (exploration) : l'utilisateur explore une zone ou un domaine particulier jusqu'à en épuiser les ressources ;
- *wandering* (errance) : l'utilisateur suit un parcours déstructuré et sans but précis.

Comme le note [Bidel *et al.* 2003], il n'y a pas de consensus général parmi les différentes recherches sur une typologie des stratégies de navigation, et chaque auteur est amené, en fonction du matériau d'expérimentation sur lequel il base son étude, à proposer des catégories différentes. Pour autant, la grande majorité des études mettent leurs participants dans la situation de chercher dans un site donné ou sur le Web pour répondre à une question précise : cette situation d'expérimentation, quasi-prototypique pour avoir été répétée chez les uns et les autres, réduit drastiquement la réalité de la navigation sur le Web. Elle enferme contenus proposés, modes d'accès et activité de navigation dans le paradigme de la recherche d'information : ce faisant, elle conclut à des équivalences entre motifs de navigation, tâche et motivation de l'utilisateur qui sont à nos yeux abusives et réductrices. Dès lors, nous laisserons volontiers de côté ces approches orientées modélisation pour notre analyse. À cela nous opposons une approche descriptive qui s'attache à replacer les modes de navigation dans le cadre de pratique avérées, et à prendre en compte la singularité des situations, des contenus et des individus.

Web Usage Mining et analyse de *logs*

Aux côtés des recherches orientées vers la modélisation des utilisateurs, le champ du *Web Usage Mining* a vu se développer un courant plutôt centré sur l'analyse de données de trafic proprement dites. Ces travaux sont massivement centrés-serveur :

exception faite des quatre études que nous avons déjà évoquées au Chapitre 1[1], l'ensemble de travaux d'analyse de *logs* de navigation porte presque systématiquement sur des traces recueillies au niveau des serveurs Web. Nous renvoyons à la lecture du compte-rendu du Workshop, *WEBKDD'99: Workshop on Web Usage Analysis and User Profiling* ([Masand & Spiliopoulou 2000]) et des panoramas proposé dans « Web Mining Research: a survey » de R. Kosala et H. Blockeel ([Kosala & Blockeel 2000]) et dans « Web Mining – Accomplishments & Future Directions » ([Srivastava *et al.* 2003]) pour une vue assez complète et relativement récente des recherches menées dans ce cadre.

De manière générale, les méthodes utilisées font fortement appel aux outils d'analyse statistiques et à la théorie des graphes ; on se reportera volontiers à [Roddick & Spiliopoulou 2002] pour un panorama des méthodes de fouille de données appliquées à l'analyse des données temporelles. On trouve dans ce champ un nombre important d'études ; nous ne les détaillerons pas dans leur ensemble, mais citerons trois travaux qui nous semblent particulièrement intéressants et représentatifs :

- *HPG (Hypertext Probabilistic Grammar)* : Borges, Levene (en particulier [Borges & Levene 1998], [Levene & Loizou 1999] et [Borges & Levene 2000]). Pour Borges et Levene, le but est de proposer des techniques permettant d'identifier des *web trails*, c'est-à-dire des séquences de liens suivis par l'utilisateur. Pour cela, le site Web étudié est modélisé comme une « grammaire régulière » (*regular grammar*) dont les états correspondent aux pages Web et la production de règles aux hyperliens. Les sessions de navigation sont incorporées dans ce modèle afin de construire une *Hypertext Probabilistic Grammar* (HPG) à laquelle on peut appliquer des techniques de *data mining*. Ces techniques sont appliquées à des *logs* côté serveur, et ont surtout pour but d'aider les webmestres à améliorer leurs sites. Le point crucial de leur recherche est l'établissement de règles reflétant des régularités dans la navigation. Pour cela, des heuristiques « à grain fin » sont développées pour trouver l'accord entre la justesse des règles et leur nombre. Dans ce travail, les chaînes de Markov sont utilisées pour l'analyse et la prédiction.
- *WebMiner* : Cooley, Mobasher, Srivastava (en particulier [Cooley *et al.* 1997], [Cooley *et al.* 1999a] et [Mobasher *et al.* 2000]). Les travaux de ces chercheurs sont centrés sur l'application des techniques de *data mining* aux usages du Web. Dans [Cooley *et al.* 1999a], ils présentent le système WebMiner qui inclut la préparation des données pour l'analyse et met en œuvre ces techniques afin de modéliser le parcours de l'utilisateur. Dans [Cooley *et al.* 1999b], R. Cooley propose le système *WebSIFT* qui utilise le contenu et la structure d'un site pour identifier les résultats potentiellement intéressants de Web Usage Mining. Dans [Mobasher *et al.* 2000], B. Mobasher poursuit ces travaux dans l'objectif d'une personnalisation des

[1] Il s'agit de [Catledge & Pitkow 1995], [Cunha *et al.* 1995], [Tauscher & Greenberg 1997a] et [Cockburn & McKenzie 2000].

sites, c'est-à-dire d'une adaptation des contenus renvoyés en fonction des chemins suivis sur un site.

- *WUM (Web Usage Miner)* : Spiliopoulou, Faulstich et Winkler. Dans [Spiliopoulou *et al.* 1999], ces trois chercheurs présentent un outil, Web Usage Miner, capable d'agréger sous forme d'arbre les différents chemins suivis au sein d'un site, de la page d'entrée à la dernière page visitée. L'ensemble des données est stocké dans un format qui permet de les interroger *via* un langage proche du SQL, MINT. Il est ainsi possible de calculer, sous forme de requête, la probabilité qu'un utilisateur voie telle page, à partir d'un parcours ayant traversé telle ou telle page, à telle ou telle position, l'éventail des combinaisons se révélant illimité. L'outil est également pourvu d'une interface graphique permettant de visualiser les parcours et les résultats de requêtes sous forme d'arbres.

Ces approches sont intéressantes en ce qu'elles traitent réellement l'aspect séquentiel et parfois temporel des parcours (pondération par la durée passée sur la page). Toutefois, les outils et méthodes centrés-serveur ne peuvent pas être directement transposés à l'analyse de données centrées-utilisateur, pour deux raisons principales. La première tient à la redondance nécessaire dans les données : dans des *logs* de serveurs, les différentes URL sont vues un nombre assez important de fois pour observer des régularités, tandis que seule une minorité de pages et de sites sont vus dans plus d'une session pour un utilisateur donné. Le second problème tient à ce que l'approche centrée-serveur part du principe que le contenu des pages naviguées est connu, et appliquent des méthodes d'analyse statistique sur des séries de symboles représentant les pages. L'interprétation des motifs de navigation devient dès lors, en dehors de toute qualification de contenu, quasi-impossible sitôt que l'on passe du côté de l'utilisateur.

Certains travaux s'efforcent cependant d'inclure cette dimension dans l'analyse : ainsi, Acharyya et Ghosh ajoutent à l'analyse statistique « classique » des *logs* une information de « concept » rattachée à chaque page ([Acharyya & Ghosh 2003]). L'objectif est ici de prendre en compte un « changement de centre d'intérêt » de l'utilisateur au cours de la session, et de pré-segmenter les sessions sur la base des contenus visités ainsi que de mieux prédire les liens qui seront suivis en fonction de la position dans l' « arbre de concepts » (*concept tree*). Les auteurs notent une augmentation significative du taux de prédiction en ayant recours à cette méthode.

Dans la même optique, Heer et Chi présentent dans [Heer & Chi 2002] une approche de la navigation *site-centric* incluant données d'usage (*logs* du serveur), de contenu (contenu textuel des pages) et de topologie (structure de liens entre les pages) des sites. L'analyse combinée de ces trois sources de données est appliquée dans un premier temps aux données recueillies auprès d'un échantillon de 21 volontaires à qui les auteurs ont demandé d'effectuer une liste de tâches sur le site Web de Xerox ; dans un second temps, Heer et Chi utilisent les *logs* entiers du serveur lui-même sur une journée. La première étape permet de régler les pondérations correctes pour chacune des trois modalités descriptives ; la seconde conduit à la classification des sessions sur la base du contenu (poids : 0,75) et des liens (poids : 0,25) des pages. Neuf classes sont générées, qui représentent les thèmes et associations de pages typiquement visités sur le site de Xerox : achat en ligne,

support technique, catalogue des produits, etc. Il est intéressant de noter que la classe la plus importante (42% des sessions) est relative à la page d'accueil, ce que les auteurs interprètent comme le reflet d'une navigation repassant fréquemment par cette page.

Ces deux études nous intéressent directement : elles ouvrent le chemin d'un croisement entre contenus visités et formes de parcours, même si nous ne souscrivons pas à la méthode des *concept trees* employée dans [Acharyya & Ghosh 2003]. Pour autant, elles restent tributaires de l'approche côté serveur : comme l'a montré [Padmanabhan *et al.* 2001] en travaillant au niveau intermédiaire d'un fournisseur de contenus disposant de *logs* relatifs à plusieurs sites commerciaux, le point de vue server-centric est partiel et biaisé, et les conclusions que l'on peut tirer de ce type d'approches sont toujours à considérer avec prudence, en particulier lorsqu'elles tendent à dresser des utilisations-types et des comportements de navigation généraux. Dans le champ du commerce électronique, il a ainsi été montré dans [Licoppe *et al.* 2002], basé sur des données de trafic centrées-utilisateur et sur des entretiens avec des internautes, que les consommateurs en ligne ont un comportement très volatil, oscillant entre achat réfléchi et achat d'impulsion, et que l'achat en ligne implique la mobilisation de ressources hors des sites de e-commerce (moteurs, comparateurs, etc.) et hors Web. Ce retour de la sociologie des usages sur l'étude de la navigation et des usages du Web montre, s'il était nécessaire, la nécessité de se placer résolument du côté de l'utilisateur et de la complexité des pratiques dans et hors Web pour appréhender les comportements de navigation.

> ***Synthèse.*** *Les travaux menés dans le champ des sciences cognitives sur la navigation dans des hypertextes ont mis à jour des motifs élémentaires de navigation qui nous intéressent directement, même si les conclusions en termes de comportements des utilisateurs ne nous satisfont pas complètement. Les recherches centrées-serveur proposent quant à elles des méthodes intéressantes d'analyse de séquences de navigation, mais leur transposition dans une approche centrée-utilisateur reste problématique.*

4.2.2 Indicateurs topologiques

À la plupart des travaux existants sur la navigation Web, nous opposons un double décalage : en premier lieu, nous adoptons une approche centrée-utilisateur qui nous amène à considérer l'ensemble des parcours sur le Web effectués par des utilisateurs identifiés. D'autre part, notre approche vise la description et non la modélisation, et se situe résolument dans le cadre des sciences humaines. Nous ne rejetons pas les méthodologies statistiques élaborées qui ont pu être développées jusqu'alors dans d'autres travaux, mais elles sortent du champ de notre travail. Pour traiter la complexité et la spécificité de nos données de trafic, ainsi que la diversité des contenus et des comportements observés, nous proposons en contrepartie des indicateurs simples de la topologie et du rythme des parcours qui, combinés aux descriptions des contenus visités, permettent de rendre compte de manière compréhensive de l'activité de navigation.

Échelle d'analyse et descripteurs

Les descriptions des sessions que l'on peut construire s'appuient sur les éléments minimaux qui les composent : requêtes HTTP, composant des pages, regroupées au sein de sites. Il importe à ce niveau d'analyse de voir ce que l'on retient de ces descriptions élémentaires et de la manière de les combiner.

Comme nous l'avons déjà souligné, l'URL ne correspond pas systématiquement à la page en tant qu'unité ergonomique. Pour les analyses centrées-serveur, ceci ne pose pas de problème insurmontable : il est possible de corriger localement les *logs* des serveurs pour tenir compte des différents systèmes de publications et modes d'organisation de chaque site. Pour l'analyse centrée-utilisateur, il en va tout autrement, car nous ne savons pas, pour chaque site, comment celui-ci est organisé et si la correspondance entre page et requête (donc URL) est juste ; l'exemple de session donné dans le Tableau 4.1 illustre ce phénomène.

Tableau 4.1. Session à plat au niveau des URL (SN2002 - session 435)

Site	Date	URL
Wanadoo (fr)	16:34:53	http://www.wanadoo.fr/bin/frame.cgi
	16:34:55	http://www.wanadoo.fr/personnalisation/bin/webauth_aff.cgi
	16:34:55	http://www.wanadoo.fr/common/abonnes/menu_accueil.html
monster.fr	16:36:10	http://www.monster.fr/
	16:36:51	http://offres.monster.fr/
	16:36:53	http://offres.monster.fr/
	16:39:42	http://offres.monster.fr/getjob.asp?JobID=14153005&col=&cy=&brd=&lid=&fn=&q=&AVSDM
Google	16:40:32	http://www.google.fr/
	16:40:43	http://www..google.fr/search?q=Hachette+Livre+%28Edition%29+&hl=fr&btnG=Recherche+G(…)
	16:41:42	http://www.google.fr/search?q=Hachette+Livre+(Edition)+&hl=fr&cr=countryFR&start=10&sa=N
hachette.com	16:42:51	http://www.hachette.com/HomePageFO/francais/site/index.htm
	16:42:52	http://www.hachette.com/HomePageFO/francais/site/blanc.htm
	16:42:53	http://www.hachette.com/HomePageFO/francais/site/blanc.htm
	16:42:53	http://www.hachette.com/HomePageFO/francais/site/blanc.htm
	16:42:53	http://www.hachette.com/HomePageFO/servlet/CtlHome?URL=site/myi-home.jsp
	16:42:54	http://www.hachette.com/HomePageFO/francais/site/blanc.htm
	16:42:54	http://www.hachette.com/HomePageFO/servlet/CtlHome?URL=site/myi-home.jsp
	16:42:54	http://www.hachette.com/HomePageFO/francais/site/blanc.htm
	16:44:26	http://www.hachette.com/HomePageFO/francais/site/page/Frameset_Groupe.jsp?page=carrieres
	16:44:29	http://www.hachette.com/HomePageFO/francais/site/page/NavHaut.htm
	16:44:30	http://www.hachette.com/HomePageFO/francais/site/page/NavGauche_Groupe.jsp
	16:44:31	http://www.hachette.com/HomePageFO/francais/site/page/Frame_Carrieres.jsp?rub=metier1
	16:44:32	http://www.hachette.com/HomePageFO/francais/site/page/NavHautInter_Carrieres.jsp?Nrubrique=1
	16:44:33	http://www.hachette.com/HomePageFO/francais/site/page/NavHaut.htm
	16:44:35	http://www.hachette.com/HomePageFO/francais/site/CAR/CAR01_ACCUEI_F.htm
	16:45:06	http://www.hachette.com/HomePageFO/francais/site/CAR/CAR06_COMMER_F.htm
	16:45:18	http://www.hachette.com/HomePageFO/servlet/CtlOffres?ACTION=0
	16:45:28	http://www.hachette.com/HomePageFO/francais/site/OFF/OFF04_STAGES_F.jsp
	16:45:29	http://www.hachette.com/HomePageFO/servlet/CtlOffres?ACTION=0
	16:45:32	http://www.hachette.com/HomePageFO/servlet/CtlOffres?ACTION=0
	16:45:44	http://www.hachette.com/HomePageFO/francais/site/CAR/CAR01_ACCUEI_F.htm

Dans cette session envisagée à l'échelle de l'URL, on compte 31 requêtes passées par l'internaute pour 26 URL distinctes, certaines étant envoyées deux fois. Les trois requêtes introductives envoyées à Wanadoo correspondent à une page unique, la page d'accueil du portail, que l'internaute a sans doute conservée en page de démarrage de son navigateur. L'analyse au niveau de la page se trouve brouillée par des données orientées trafic et peu fiables à l'échelle de la page : impossible, dès lors, de comparer les visites de sites en nombre de pages vues, car chaque webmestre aura mis en place des systèmes de *frames* différents, ce qui rend les résultats non homogènes.

Au-delà de ce problème, on questionnera volontiers l'approche « nombre de pages », qui a été beaucoup employée dans la mesure d'audience à ses débuts : quand bien même les données de trafic nous permettraient de compter exactement les pages vues du point de vue de l'utilisateur, quel sens donner à un tel décompte ? Certes, avoir ou ne pas avoir visité une page, voilà déjà un indice de fréquentation indéniable ; mais si l'on veut aller plus loin et entrer dans une logique comparative d'analyse de la fréquentation des sites Web, le décompte des pages soulève plus de problèmes qu'il n'en résout. Les pages, nous avons eu l'occasion de le constater dans l'examen préliminaire du corpus constitué à partir des données BibUsages, diffèrent en termes de taille, de fonction, de types de contenus : de très longues et de toutes petites, des textes volumineux et de simples formulaires, etc. De ce point de vue, la visite de la page ne prend sens que dans la dynamique de la navigation, et uniquement dans la mesure où nous pouvons en décrire le contenu thématique et/ou fonctionnel. De ces éléments, nous tirons deux conclusions méthodologiques importantes.

En premier lieu, il est nécessaire d'adopter la bonne échelle d'analyse en fonction des informations dont on dispose. Nous avons déjà quelque peu abordé la question de l'échelle de représentation des parcours dans la présentation des outils de visualisation, et avions vu que des échelles différentes donnent des résultats visuels sensiblement variables. Cette question se pose de manière plus brutale encore lorsque l'on souhaite construire des représentations chiffrées des parcours : en particulier, la linéarité de la navigation diffère grandement, pour un parcours donné, selon que l'on considère les pages visitées, les sites accédés, les services utilisés.

Dans l'exemple donné au Tableau 4.1, nous avions observé que la session est non linéaire au niveau de la page ; toutefois, si l'on observe cette session au niveau du site, elle est strictement linéaire, l'internaute ne revenant pas sur un site déjà vu au cours de la session. La session est alors décomposable en quatre pas : 1) Wanadoo (fr), 2) monster.fr, 3) Google et 4) hachette.com. On notera à cet instant les bénéfices du préformatage et de l'enrichissement des données brutes : d'une part, les domaines `www.monster.fr` et `offres.monster.fr` sont regroupées en un seul et même site, 'monster.fr' ce qui est cohérent avec une logique centrée-utilisateur ; d'autre part, l'identification des portails généralistes avec *CatService* permet de repérer que les pages vues sur `www.wanadoo.fr` sont relatives au portail Wanadoo (fr), taggué comme 'Portail généraliste', et que celles sur `www.google.fr` se rapportent à Google, 'Moteur de recherche'.

On peut également exploiter plus finement les descriptions fournies par *CatService* (voir Tableau 4.2). On sait alors que les trois URL vues sur Wanadoo correspondent à la page d'accueil, tandis que sur Google, le mouvement est décomposable en 1) l'accès à la page d'accueil (une URL), et 2) une requête contenant les mots-clefs « Hachette Livre (Edition) » (deux URL).

Tableau 4.2. Session agrégée au niveau du site et des services (SN2002 - session 435)

Date	site/portail	Nb URL	Durée	Service	mots-clefs
16:34:53	Wanadoo (fr)	3	7"	Page Accueil	
16:36:10	monster.fr	4	4' 12"		
16:40:32	Google	1	11"	Page Accueil	
16:40:43	Google	2	2' 8"	moteur	Hachette Livre (Edition)
16:42:51	hachette.com	21	2' 54"		

En fonction des différentes sessions et des différentes échelles d'analyse, les résultats seront sensiblement différents en termes de linéarité. Devant le défaut de fiabilité des données au niveau de la page, on s'attachera plutôt par la suite à travailler au niveau du site (ou du portail), et du service lorsque celui-ci est identifié dans *CatService* ; ceci est par ailleurs cohérent avec la mobilisation des descriptions des annuaires, qui décrivent majoritairement les données de trafic à l'échelle du site.

Deuxième élément méthodologique, hors de toute qualification des contenus, on préfèrera s'attacher à la durée de visite plutôt qu'au nombre d'URL visitées. Dans l'exemple de session ci-dessus, on constate ainsi que si l'internaute a demandé quatre URL sur `monster.fr` contre 21 sur `hachette.com`, il a passé plus de quatre minutes sur le premier site, contre moins de trois minutes sur le second. Bien évidemment, rien ne nous dit qu'il ne s'agit pas là d'un effet de feuilletage plus rapide des contenus proposés par `hachette.com` par rapport à ceux de `monster.fr`. Quoi qu'il en soit, le problème est insoluble dès lors que l'on ne dispose pas d'information précise sur les contenus visités.

Dans cette perspective, si l'on agrège les données de navigation à l'échelle des sites, la durée apparaît comme une donnée préférable au nombre de pages. Certes, cet indicateur n'est pas absolu : il faut en particulier se garder de conclure à une équivalence entre durée et importance dans la navigation ou intérêt de l'utilisateur. Certains sites comme les moteurs de recherche peuvent fonctionner comme des lieux de passage où l'utilisateur ne va pas rester longtemps : il ne faut pas en conclure pour autant que leur présence est négligeable dans le parcours. Pour autant, la durée est, dans les données dont nous disposons, le moins mauvais indicateur ; c'est au moment de l'interprétation des résultats qu'il importera d'y prêter une attention particulière.

Construction d'indicateurs topologiques simples

La dimension temporelle est un des éléments fondamentaux de la sémantique des parcours. Elle se situe à deux niveaux : d'un côté, il s'agit de prendre en compte les durées de visites et le temps passé sur chaque page ou chaque site, ce que nous venons de voir. De l'autre, il importe d'examiner l'ordre dans lequel les contenus sont accédés et la valeur qu'ils prennent dans la dynamique du parcours.

Pour cela, nous avons tenté de nous munir d'outils de fouille permettant de vérifier ces hypothèses, ainsi que de mettre en place des outils statistiques à même de rendre compte de la dynamique des parcours. L'approche que nous mettons en place est simple, et repose sur la construction d'indicateurs permettant de représenter les « formes » et le rythme des parcours de manière synthétique. Les indicateurs doivent permettre de représenter certains aspects particuliers de la session :

- s'il est linéaire ou non ;
- le nombre et la longueur des détours ;
- l'importance de ces détours dans la temporalité de la session ;
- distinguer et quantifier les points de fixation de la session (centres de formes en rosace).

Le parti pris de la simplicité et de la robustesse qui est le nôtre nous amène à n'envisager que partiellement la dimension séquentielle : seul un outillage complexe permettrait de tenir ensemble, dans un même objet statistique, les éléments de forme et de contenus des parcours Web[1]. Dans les indicateurs que nous proposons, la séquentialité est analysée en dehors des éléments de contenu : nous formalisons les sessions Web comme une séquence de symboles représentant les éléments visités. À partir de cette représentation, nous construisons les indicateurs suivants :

- N : longueur de la session (nombre de pas) ;
- n : nombre d'éléments uniques vus dans la session ;
- $r = \frac{n}{N}$: taux moyen de linéarité du parcours, qui vaut 1 s'il est linéaire et se rapproche de 0 au fur et à mesure que cette linéarité diminue ;
- R : nombre d'éléments revisités, c'est-à-dire vus plus d'une fois ;
- $c = \frac{N-n}{R}$: nombre moyen de revisites par élément revisité. Cet indicateur représente la concentration des revisites sur un ou plusieurs éléments du parcours, et permet de détecter des navigations « en étoile » : dans l'exemple de la Figure 4.8, pour les deux navigations N=12, n=9, et r=1,3, mais c est différent (3 vs. 1).

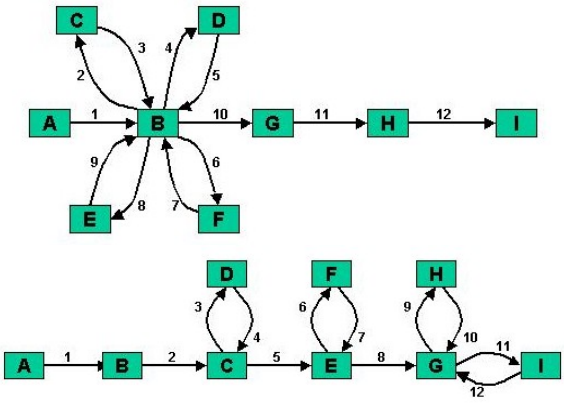

Figure 4.8. Concentration des revisites

[1] Nous pensons en particulier aux graphes colorés, qui permettent d'attacher des informations aux nœuds et aux arcs du graphe, tout en tenant compte du caractère orienté et temporel de l'objet.

Nous construisons également des indices qui prennent en compte les durées passées sur chaque élément visité, tout en tenant compte des séquences de navigation :

- *T* : durée totale de la session ;
- durées moyenne et médiane passées sur chaque pas de la session ;
- *T1* : le temps passé sur les éléments de la session vus une seule fois ;
- $d = \frac{T1}{T}$: part du temps passé sur des éléments vus une fois dans l'ensemble de la session. Cet indicateur est proche du taux de linéarité *r*, mais s'applique aux durées : il vaut 1 si la session est linéaire, et 0 si elle ne l'est pas du tout.

Nous avons également souhaité avoir des informations qualitatives sur la façon dont les pages sont revisitées. En particulier, nous avons voulu mesurer l'emploi de la fonction *Back* des navigateurs (retour d'une page en arrière). Pour cela, nous avons développé un algorithme spécifique capable d'identifier les séquences de *back* et de les isoler du reste de la session. Pour chaque session, nous en générons une nouvelle représentation qui correspond au parcours sans les mouvements de *back*. Par exemple, une session "A→B→C→D→E→D→C→F" sera transformée en "A→B→C→F", la séquence "C→D→E→D→C" étant réduite à "C". La Figure 4.9 illustre ce travail de réécriture des sessions, et montre en particulier la différence entre les séquences de type *back* et les boucles, non exclues dans ce traitement.

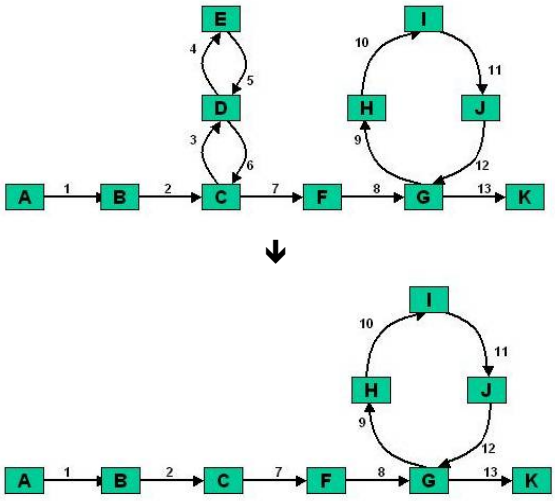

Figure 4.9. Réécriture d'une session sans les mouvements de Back

Ainsi, nous produisons une nouvelle série d'indicateurs relatifs à l'utilisation du *Back* et aux sessions dont les *back* ont été ôtés :

- *B* : nombre de séquences de type *Back*, quelle que soit leur longueur ;
- N_b : la longueur du parcours (nombre de pas) une fois les séquences de *back* ôtées ;

- $b = \frac{N_b}{N}$: part des actions de type *Back* dans le nombre total de pas dans la session. Plus l'indice est proche de 0, plus les actions *Back* occupent de place dans la session.

La quantification des actions de type *Back* est intéressante à double titre. À l'échelle de la page, elle correspond à l'utilisation d'une fonctionnalité des navigateurs, et rend compte d'un mode d'utilisation des interfaces et du déroulement du parcours. Au niveau du site, la correspondance avec la fonctionnalité des navigateurs n'opère que si une seule page est vue sur chaque site de la séquence d'aller-retour, et ne renvoie donc pas tant à une fonctionnalité de l'IHM qu'à renforcer l'identification de sites-pivots au sein d'une navigation en étoile.

Comme nous l'avons discuté précédemment, la session peut être abordée à deux niveaux de complexité : page / site, et nombre d'éléments / durée. En conséquence, chacun des indicateurs décrits ci-dessus est dédoublé, selon l'échelle retenue. La liste d'indicateurs finale obtenue est donnée au Tableau 4.3 ci-dessous.

Tableau 4.3. Liste des indicateurs topologiques et temporels retenus

	Indicateur	Description
	D	Durée de la session (en secondes)
Échelle : page	P	Nombre de pages visités (nombre de pas)
	p	Nombre de pages distinctes visitées
	R_{page}	Nombre de pages distinctes vues plus d'une fois
	r_{page}	Taux de linéarité – échelle page
	c_{page}	Taux de concentration – échelle page
	B_{page}	Nombre d'action de type *Back* – échelle page
	b_{page}	Part des actions de type *Back* dans la session – échelle page
	t_{page}-*moy*	Durée moyenne sur chaque page vue
	t_{page}-*med*	Durée médiane sur chaque page vue
	$D1_{page}$	Durée totale sur les pages vues une seule fois
	d_{page}	Part du temps passé sur les pages vues une seule fois
Échelle : site	S	Nombre de sites visités (nombre de pas)
	s	Nombre de sites différents visités
	R_{site}	Nombre de sites distincts vus plus d'une fois
	r_{site}	Taux de linéarité – échelle site
	c_{site}	Taux de concentration – échelle site
	B_{site}	Nombre d'action de type *Back* – échelle site
	b_{site}	Part des actions de type *Back* dans la session – échelle site
	t_{site}-*moy*	Durée moyenne sur chaque site vu
	t_{site}-*med*	Durée médiane sur chaque site vu
	$D1_{site}$	Durée totale sur les sites vus une seule fois
	d_{site}	Part du temps passé sur les sites vus une seule fois

Si ces indicateurs simples ne rendent pas compte de la complexité des formes de sessions dans son ensemble, ils en donnent un bon aperçu. Ils permettent d'établir des premières segmentations élémentaires des sessions sur la base de leur topologie ; couplés aux outils d'examen manuel des parcours et aux descripteurs de contenu, ils

doivent permettre de croiser forme et contenu de parcours et de parvenir à une segmentation des activités de navigation.

> ***Synthèse.*** *Nous avons élaboré des indicateurs statistiques simples pour représenter la topologie et la temporalité des parcours, à l'échelle de la page comme du site : linéarité, concentration des revisites, temps passé sur les pages et les sites dans et hors des retours, utilisation de la fonction* back *sont représentés de manière synthétique. Ces indicateurs alimentent une analyse praxéologique de la navigation.*

4.3 Contextualisation

La description de la sémantique des parcours que nous élaborons place la session au cœur de l'analyse, et en fait son objet privilégié. Gardons toutefois à l'esprit que dans l'activité de navigation comme ailleurs, le global définit le local : dans cette perspective, on s'efforcera autant que possible de replacer les parcours dans le double contexte de l'offre de contenu et du profil de l'utilisateur.

4.3.1 Contexte global du Web

Au niveau macro-analytique, avant d'avancer dans l'analyse des parcours sur le Web de page en page et de site en site, il semble important d'avoir une connaissance de l'arrière-plan structurel du Web. Un certain nombre de travaux ont été menés sur la structure du Web, modélisant celui-ci comme un graphe dont les noeuds sont les pages et les arcs les liens d'une page vers l'autre, mais on a vu peu d'études systématiques avec des robots parcourant l'ensemble du Web. Parmi ces dernières, une certaine controverse a semblé exister entre les résultats des différentes équipes ; le travail de Broder *et alii*, présenté en 2000 ([Broder *et al.* 2000]) s'impose dans les débats et apparaît comme le plus exhaustif et le plus fiable de tous.

Le résultat de ces analyses, dont nous reproduisons la représentation graphique (Figure 4.10 ci-dessous), montre une forme en « nœud papillon » :

- au centre, un réseau très fortement interconnecté, qu'il est possible de parcourir facilement ;
- à gauche, des sites qui pointent vers la nébuleuse centrale mais qu'il est difficile de rejoindre car peu de liens permettent de s'y rendre, typiquement des sites personnels à faible notoriété ;
- à droite, au contraire, une série de sites désignés par les pages du groupe central mais dont il est difficile de sortir car ils renvoient peu ou pas vers d'autres sites : ce groupe est essentiellement composé de sites commerciaux, qui contiennent peu de liens externes ;
- quelques passages directs de la partie gauche à la partie droite qui « évitent » la partie fortement interconnectée du Web.
- enfin, de petits composants fortement interconnectés mais sans liens avec le reste du Web.

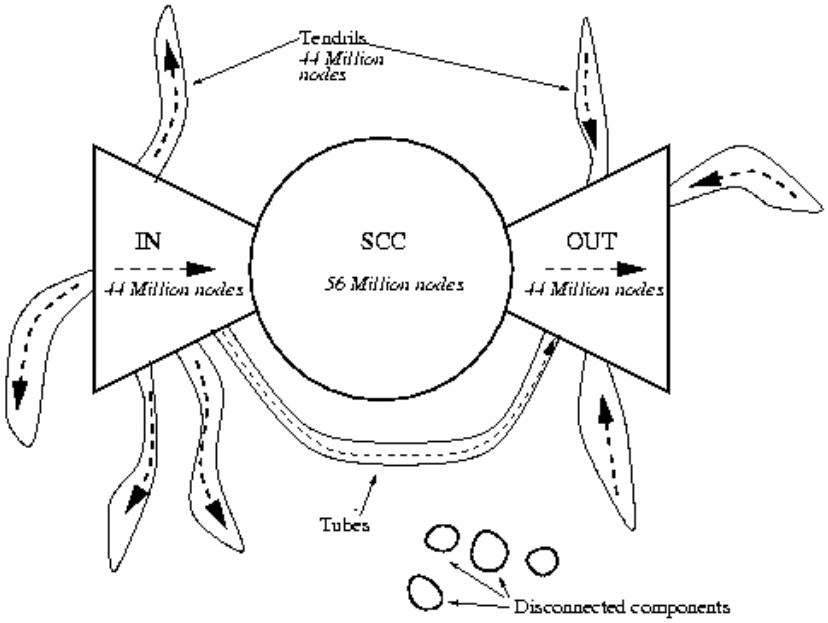

Figure 4.10. Structure du Web représentée dans [Broder et al. 2000]

Ce travail se trouve prolongé dans [Baeza-Yates & Poblete 2003], qui offre un éclairage longitudinal sur l'évolution du Web chilien entre 2000 et 2002. S'appuyant sur les travaux de Broder, l'auteur montre que la croissance constante de la Toile chilienne s'accompagne d'une complexification des sites et d'un accroissement du « noyau central » fortement interconnecté.

Dans le même champ d'analyse, on remarquera également le travail présenté dans [Faloutsos *et al.* 1999], qui souligne la pertinence des lois statistiques de type loi de puissance, « power-law » (i.e. de Pareto) pour l'analyse de la topologie du Web ; [Adamic 1999], [Adamic & Huberman 2001] et [Adamic 2001], où Adamic observe que le « Small World Web » est un monde de taille restreinte, au sens ou les pages et les sites sont regroupés en petits réseaux fortement interconnectés et de nature communautaire ; et enfin [Maurer *et al.* 2000] qui analyse la structure particulière que constituent les *Web rings*.

Ces observations sont particulièrement intéressantes : dans une navigation de site en site, l'impossibilité de sortir d'un site ou au contraire l'impossibilité d'aller sur un site car aucun lien n'y mène sont des facteurs importants. Il serait par exemple très utile de savoir quelle est la position de chaque site visité par un internaute sur le graphe de Broder. Si cette ambition dépasse le cadre de notre travail, notons ici l'intérêt d'un tel projet.

> ***Synthèse.*** *L'analyse macroscopique du Web sous l'angle des liens hypertextes, des flux de navigation et des cliques de sites peut apporter un arrière-plan pertinent pour l'analyse des parcours. Cette piste, complexe à mettre en œuvre, ne sera pas exploitée ici, mais on gardera à l'esprit ces éléments macro-structurels sous-jacents à la navigation.*

4.3.2 Contexte de l'utilisateur

Pratiques Web, pratiques Internet

Un premier élément de contextualisation de la navigation au niveau de l'individu consiste à replacer chacune de ses sessions dans le corpus global de ses parcours sur le Web. Ce type d'étude, qui nécessite une durée d'observation suffisamment longue pour être probante, est aussi rare que le sont les travaux centrés-utilisateur : dans [Catledge & Pitkow 1995] (trois semaines d'observation, 107 utilisateurs) ou [Crovella & Bestavros 1996] (cinq mois et demi, pour 37 postes équipés du dispositif de recueil de trafic), les sessions sont considérées en masse et ne sont pas rapportées à tel ou tel utilisateur.

Par contre, dans [Tauscher & Greenberg 1997a] et [Tauscher & Greenberg 1997b] (23 utilisateurs durant six semaines), les auteurs se penchent spécifiquement sur la « revisite » de pages Web, et s'attachent à examiner pour chaque internaute la fréquence de visite des pages accédées ; l'analyse montre qu'en moyenne 58 % des requêtes pour un utilisateur donné pointent vers une page qu'il a déjà visitée, en même temps que le « vocabulaire » des URL ne cesse de croître avec le temps. L'étude présentée dans [Cockburn & McKenzie 2000] (70 utilisateurs observés pendant quatre mois) va dans le même sens : pour chaque internaute observé, Cockburn et McKenzie constatent une croissance assez régulière du vocabulaire (les pages) dans le temps, une forte corrélation entre le nombre de visites et le vocabulaire, ainsi qu'une distribution de type zipfienne de la visite des pages (peu de pages sont visitées très régulièrement, beaucoup le sont rarement).

Ces premiers constats statistiques conduisent d'ores et déjà à considérer que la navigation amène sans cesse à voir de nouveaux sites, mais à n'en revoir que très peu : dans le corpus de pages et de sites de chaque internaute, on peut distinguer de manière nette les sites vus une fois seulement et ceux soumis à des visites routinières. Routinier ne signifie pas forcément fréquent : un internaute peut être amené à consulter systématiquement le ou les mêmes sites dans un contexte donné, tout en ne se trouvant que rarement dans cette situation précise : préparer ses vacances en allant sur des sites de voyagistes, réserver un billet de train, connaître un itinéraire routier, chercher un emploi sont autant d'activités qui, sans être fréquentes, peuvent être contextuellement régulières.

Nous souhaitons approfondir ce point : derrière la notion de corpus de sessions et de sites pour chaque individu, se profile la question de la construction des territoires personnels sur le Web. Découverte de sites, évolution des modes de navigation, modes d'appréhension des sites en fonction de leur place dans les routines de l'utilisateur, apprentissage ou perfectionnement dans le maniement des outils Web sont autant de questions qui renvoient à une éthologie de la navigation et un examen des modalités de la pratique en contexte. En outre, l'offre étant en perpétuel renouvellement, tant au niveau des sites que des types de contenus et de services proposés (« webisation » des outils de communication, outils de publication simplifiés, élargissement de l'offre de contenus culturels, commerciaux, etc.), l'internaute est soumis à la nécessité d'adapter ses comportements à cette évolution et se trouve potentiellement en perpétuelle situation d'apprentissage et de découverte. Nos données de trafic nous invitent à ce type d'approche : centrées-

utilisateur et exceptionnelles par leur taille et leur durée (plusieurs milliers d'internautes suivis durant plusieurs mois ou plusieurs années), elles ouvrent légitimement la voie d'une approche longitudinale approfondie.

Un autre élément de contextualisation individuelle des pratiques tient à l'examen de l'activité Internet dans son ensemble. Il s'agit là d'un élément fondamental dans l'approche retenue pour les projets TypWeb, SensNet ou BibUsages : la navigation sur la Toile s'insère et s'entrelace avec les autres outils Internet, comme la messagerie électronique, le *chat*, les jeux en ligne, le téléchargement, etc.

Dans le cadre du projet TypWeb, V. Beaudouin a établi une segmentation d'une cohorte d'internautes résidentiels fondée sur l'utilisation du Web et des outils de communication :

> Deux grands groupes d'internautes se distinguent : ceux qui accordent une place prépondérante au Web dans leurs usages d'Internet et ceux qui favorisent au contraire l'usage des services de communication. Dans chacun de ces groupes se constituent des axes de différenciation, en fonction de l'intensité d'usage pour le premier groupe et en fonction du ou des outils de communication utilisés (mail classique, WebMail, *chat*, messagerie instantanée…) pour le second groupe. Les utilisateurs de *chat* et messageries instantanées se recrutent surtout chez les jeunes, et se distinguent par leur capacité à articuler au cours d'une même session consultation du Web, utilisation de la messagerie et conversations synchrones.[1]

Dans une étude sur l'entrelacement des médias dans la constitution des publics de l'émission Loft Story (voir [Beaudouin *et al.* 2003a]), nous avions également constaté la synchronisation de l'activité Internet avec le flux télévisuel et la création de communautés sur support électronique extrêmement actives mobilisant le Web comme support de publication, les salons de *chat* comme espace de débat collectif et d'échanges inter-individuels, et la messagerie comme outil d'échange d'images ou d'informations.

Ces éléments rappellent que l'activité de navigation s'inscrit dans un double entour : celui des contenus et des services Internet en général, avec lesquels elle se trouve étroitement entrelacée, et l'entour plus global des pratiques hors Internet. Si ce deuxième point nous échappe, nous prendrons en compte autant que faire se peut l'usage global d'Internet dans la description des utilisateurs, afin de voir dans quelle mesure cette dimension générale permet d'expliquer des comportements de navigation sur le Web. Gardons à l'esprit que la navigation sur le Web s'apparente à une activité « comme une autre » et qu'elle s'inscrit dans l'univers des pratiques de l'individu, ce qui justifie encore, s'il était besoin, une approche praxéologique et contextualisée de la navigation.

---

[1] [Beaudouin *et al.* 2002], p. 6.

Éléments socio-démographiques

Les études menées dans le champ de la sociologie des usages nous rappellent que les pratiques relatives à Internet ne sont pas neutres socialement[1]. En dehors du constat d'une fracture numérique, qui touche en particulier la capacité à s'équiper en terminaux pour les ménages à faibles revenus, [Lelong 2003] fait remarquer qu'« il reste à en préciser les multiples dimensions qui ne se limitent pas aux inégalités d'accès aux nouvelles technologies ». L'auteur précise que « l'analyse de leurs usages permet notamment d'évaluer l'importance de l'âge, du sexe, du milieu social dans cette appropriation ou ce rejet des nouveaux outils que sont principalement l'ordinateur et Internet. ». L'utilisation et l'appropriation de l'outil informatique et, corrélativement, des outils Internet sont encore aujourd'hui très sexuées, même si les femmes investissent aujourd'hui ce terrain longtemps demeuré masculin. Comme le remarque Jouët dans [Jouët 2003] à propos des TIC en général, « les catégories binaires – technologie/homme, relation/femmes – sont plus complexes qu'il n'y paraît. On observe ainsi une inversion des qualités attribuées à chaque sexe : les femmes traditionnellement associées à la subjectivité et à l'émotion, font preuve d'une grande rationalité dans leurs usages, alors que les hommes, traditionnellement rangés du côté de l'objectivité et de la rationalité, donnent libre cours à leur émotion et à leurs affects dans leur relation à la machine »[2].

En termes de milieux sociaux, il a été montré dans le cadre du projet TypWeb que les usages des outils de communication sont variables selon l'âge et la catégorie socio-professionnelle de l'utilisateur (voir [Beaudouin *et al.* 2002]). En matière d'outils de communication, les internautes utilisant de façon privilégiée le courrier électronique sont plus fréquemment des cadres et professions intermédiaires ; à l'inverse, les outils de communication synchrone (*chat*, messagerie instantanée) sont préférentiellement utilisés par les jeunes et les individus appartenant à des foyers dont le chef de famille est employé ou ouvrier. L'hypothèse avancée par Beaudouin pour expliquer cette différence, relate [Lelong 2003], repose sur la distance à la culture légitime et les barrières à l'écrit propres aux jeunes adultes issus de milieux modestes : « ainsi s'expliquerait leur préférence pour des échanges écrits rapides, quasi conversationnels, sans traces durables et donc moins exposés que le mail à des jugements sociaux valorisant la correction orthographique et grammaticale, et érigeant la lettre manuscrite en modèle de communication. »

La navigation sur le Web est également influencée par ces clivages sociaux : « Dans les familles des milieux favorisés, la pratique de lecture des livres est valorisée et structurée selon les schémas de la culture humaniste et classique. […] Plus on descend dans l'échelle sociale, plus les lycéens décrivent Internet comme un gisement de connaissances digne de foi et supérieur aux autres. »[3]

---

[1] Voir [DiMaggio *et al.* 2001] pour un panorama des recherches en sociologie sur Internet, structurées autour de cinq thématiques : les inégalités (*digital divide*), les communautés et les collectifs, les implications politiques, l'impact sur les organisations, et la diversité culturelle.

[2] [Jouët 2003], p. 81.

[3] [Lelong 2003], p. 114.

Cette inscription sociale des pratiques Web n'entre que partiellement dans notre champ d'investigation : notre objectif n'est pas de différencier les pratiques sur la base des catégories socio-professionnelles, mais ces variables peuvent localement être mobilisées pour expliquer et interpréter des comportements. Elle agit également comme garde-fou : on s'attachera à tenir compte des déterminants socio-économiques dans l'examen des centres d'intérêt, des contenus des parcours, de l'expertise, tant il est vrai que le capital social, culturel et technique est un déterminant important des pratiques.

> ***Synthèse.*** *L'usage du Web n'échappe pas plus que toute autre activité à des déterminations sociales, qui influencent les parcours sur le plan des contenus autant que des modalités. Pour autant, c'est surtout dans l'usage de la Toile que l'on cherchera les éléments de contextualisation les plus pertinents. L'étude des pratiques d'Internet en général et l'analyse de la structure et des modes d'appréhension des territoires personnels alimentent ainsi une approche éthologique et contextuelle des parcours sur le Web.*

Conclusion

En positionnant l'étude des parcours sur le Web dans le champ des sciences humaines et sociales, nous nous écartons sensiblement des travaux qui ont pu être menés jusqu'alors sur la navigation : tentatives de modélisation des comportements d'utilisateurs fortement influencées par les sciences cognitives d'une part, et point de vue centré-serveur d'autre part. Notre démarche est descriptive, et se concentre sur l'activité de navigation du côté de l'utilisateur, avec la diversité des situations et des contenus que cela implique. Pour appréhender cette complexité, nous avons élaboré des outils complémentaires de fouille des données de trafic : de manière qualitative, la possibilité de visualiser et de refaire les parcours permet à la fois de formuler et de vérifier des hypothèses ; les indicateurs topologiques simples que nous avons mis en place autorisent quant à eux des traitements statistiques de masse sur la forme, le rythme et la temporalité des parcours. Couplés aux descriptions de contenus au niveau de la page et du site d'une part, et aux éléments de contextualisation individuelle des pratiques d'autre part, ils forment une base solide pour l'analyse des parcours, leur description, leur segmentation et leur interprétation.