II

Usages et comportements de navigation sur le Web

A partir des données brutes de trafic, nous nous sommes dotés d'un outillage complet pour l'analyse des parcours : représentation des contenus à l'aide des annuaires Web et de l'application *CatService*, indicateurs statistiques rendant compte de la forme et de la temporalité des sessions, éléments de contextualisation sous l'angle des territoires personnels et des pratiques Internet, modules de visualisation et d'examen manuel des parcours. On dispose ainsi d'une description complète des parcours intégrant, de la page à l'utilisateur, la description de la production et de l'offre sur le Web et les modalités de leur appréhension par les internautes.

Cette seconde partie montre la mise en œuvre de ces outils et descripteurs sur trois panels d'internautes centrés-utilisateur pour l'analyse des usages et des comportements de navigation sur le Web. Dans un premier temps, nous centrons notre analyse sur une vue locale et décontextualisée des parcours afin de construire une segmentation des sessions basée uniquement sur leur forme et leur contenu. Les cinq parcours-type que nous identifions sont ensuite examinés en regard des pratiques d'Internet et des territoires personnels des utilisateurs sur le Web ; cet examen contextuel de la navigation nous amène à identifier des modes prototypiques d'appréhension des sites Web en fonction de leur contenu et de leur valorisation par l'utilisateur.

Chapitre 5

Contenus et formes de parcours

Nous présentons dans ce chapitre les trois sources de données que nous avons exploitées pour cette étude : un panel résidentiel général large en 2002, une cohorte suivie à domicile de 2000 à 2002, et un échantillon restreint fin 2002 centré sur les usages des bibliothèques électroniques. Après avoir décrit la composition de ces panels, leurs spécificités et les résultats particuliers que nous souhaitons en retirer, nous donnons un portrait général des usages du Web en termes d'intensité, de régularité, de thématiques et de services qui nous amène à établir une première segmentation des parcours.

5.1 Description des panels

Nous travaillons sur trois sources de données complémentaires. La première porte sur l'année 2002 et concerne un panel large de 3 372 individus ; la seconde, longitudinale, est centrée sur 597 internautes français observés de 2000 à 2002 ; la dernière est un panel de taille plus modeste formé de 72 utilisateurs de bibliothèques électroniques.

5.1.1 Panel SensNet 2002

Le panel SensNet 2002 dont nous disposons est issu des données de trafic recueillies par NetValue dans le cadre du projet SensNet (voir Annexe 1). Cette cohorte est constituée de 3 372 internautes dont l'activité Web est observée de janvier à octobre 2002, soit 10 mois d'observation. Ce panel, désigné dans la suite du document par l'abréviation 'SN2002', a été élaboré avec la méthode des quotas par la société NetValue : il est globalement représentatif de la population des internautes français sur la période. En effet, nous ne disposons pas ici du panel de NetValue proprement dit : celui-ci, conçu pour offrir une vue représentative de la population des internautes à tout moment à des fins de mesure d'audience, est corrigé chaque mois pour y intégrer les nouveaux internautes et assurer sa représentativité. Notre échantillon suit une logique quelque peu différente : il s'agit d'étudier une cohorte

d'internautes et d'observer l'évolution des usages de ces individus sur une période donnée.

Le panel SN2002 est alors représentatif de la population des internautes français en janvier 2002, mais cette représentativité s'affaiblit au fil des mois. C'est au chapitre de l'ancienneté de la pratique que l'échantillon est le plus biaisé par rapport à la population des internautes en général (voir Tableau 5.1) : les primo-accédants sont absents des données, les internautes « récents » sous-représentés (5,5 % du panel connecté pour la première fois en 2001), la moitié du panel a deux à trois ans de pratique, et le tiers était connecté en 1998 ou avant.

Tableau 5.1. SN2002, ancienneté de la pratique Web

Année de première connexion	% du panel
Avant 1997	8,8 %
1997	7,9 %
1998	18,2 %
1999	26,8 %
2000	26,8 %
2001	5,5 %
Ne sait pas	6,0 %

Malgré cette sur-représentation des « anciens internautes », le profil du panel en termes d'âge, de sexe et de type de connexion suit globalement ceux de l'ensemble des internautes français à la même époque (voir Tableau 5.2).

Tableau 5.2. SN2002 : caractéristiques générales

		SN2002 en janvier 2002	Internautes français en décembre 2001
Homme / femme		56,4 % / 43,6 %	58 % / 42 %
Âge	Moins de 15 ans	7,8 %	4,6 %
	15-24 ans	22,8 %	24,7 %
	25-34 ans	22,3 %	23,1 %
	35-49 ans	31,5 %	31,8 %
	50-64 ans	12,6 %	12,5 %
	Plus de 65 ans	3,0 %	3,3 %
Bas débit / haut débit (autre, n.s.p)		90,1 % / 7 % (2,9 %)	91,1 % / 8,9 %

Le taux d'équipement du panel en accès Internet à haut débit est similaire à celui de l'ensemble de la population, alors que nous aurions pu nous attendre, chez ces internautes confirmés, à une part importante d'abonnements au câble ou à l'ADSL : ceci est sans doute dû au fait que c'est en 2002 que l'accès haut débit s'est diffusé largement en France, notre panel étant observé ici en janvier 2002.

En termes de catégories socio-professionnelles (CSP), le panel SN2002 est également très proche des caractéristiques de l'ensemble des internautes français à la même époque (décembre 2001) ; par rapport à la population française en général, les professions intermédiaires sont surreprésentées, au détriment des ouvriers et des agriculteurs (voir Tableau 5.3).

Tableau 5.3. SN2002 : CSP des panélistes

Occupation	Nb. panélistes	Part du panel	Internautes français, déc. 2001
Retraité(e)	186	5,5 %	6,2 %
Sans profession, au foyer	416	12,2 %	9,4 %
Temporairement sans emploi (chômage, maladie, etc.)	62	1,8 %	1,6 %
Étudiant(e)	727	21,4 %	23,3 %
Ouvrier	146	4,3 %	4,4 %
Agriculteur, pêcheur	8	0,2 %	0,3 %
Employé(e)	594	17,5 %	18,8 %
Profession intermédiaire (cadre, chef de service, chef de groupe, technicien, etc.)	851	25,0 %	24,2 %
Profession libérale	69	2,0 %	1,5 %
Dirigeant (PDG, DG, directeur, cadre sup., etc.)	282	8,3 %	8,4 %
Propriétaire d'une entreprise, artisan, commerçant, autre travailleur indépendant	49	1,4 %	1,5 %
Non renseigné	8	0,2 %	-

On retrouve là des éléments présentés maintes fois dans des études et des réflexions autour du thème de la « fracture numérique », nous n'irons donc pas plus avant sur ce sujet. Dans le cadre présent, nous nous contenterons de constater que malgré une présence forte d'anciens internautes, notre échantillon est représentatif en termes de CSP.

> ***Synthèse.*** *La taille, la représentativité et la durée d'observation du panel SensNet 2002 permettent de travailler dans la masse et d'observer des comportements récurrents entre sessions et entre individus. Ce panel sert de toile de fond à la mobilisation d'autres jeux de données moins représentatifs.*

5.1.2 Panel longitudinal 2000-2002

L'échantillon longitudinal sur lequel nous travaillons est également issu des données du panel NetValue : il s'agit d'un sous-ensemble du panel SN2002, dont on observe 597 individus présents dans le panel de 2000 à 2002 (données issues des projets TypWeb et SensNet) ; nous le désignons par la suite par l'abréviation 'SN00-02'. Le mode de sélection des internautes de cette cohorte laisse de côté les abandonnistes, et favorise plus encore que dans l'échantillon précédent les internautes anciens et « fidèles », qui ont ancré l'usage d'Internet dans leurs pratiques. On ne s'étonnera pas, dès lors, de constater que la moyenne d'âge de la cohorte SN00-02 est centrée sur la tranche 35-49 ans (voir Tableau 5.4 ci-dessous), alors que les âges sont beaucoup plus répartis pour l'ensemble des internautes français à la même période.

Tableau 5.4. Panel SN00-02 : caractéristiques générales

		SN00-02 en 2000	SN00-02 en 2002
Homme/femme		67,5 % / 32,5 %	
Lieu d'habitation	Rural	16,7 %	15,2 %
	2 000 à 19 999 hab.	13,4 %	14,4 %
	20 000 à 100 000 hab.	10,2 %	10,4 %
	100 000 et + hab.	38,6 %	37,8 %
	Région parisienne	21,1 %	22,3 %
Âge	Moins de 15 ans	10,4 %	3,6 %
	15-24 ans	21,5 %	24,3 %
	25-34 ans	18,2 %	19,3 %
	35-49 ans	35,5 %	36,0 %
	50-64 ans	13,2 %	14,5 %
	Plus de 65 ans	1,3 %	2,3 %
Bas débit / haut débit (n.s.p)		nc.	89 % / 9 % (2 %)

Corrélativement, les catégories socio-professionnelles sont plus proches de celles des internautes français avant 2000 qu'en 2002 (voir Tableau 5.5) : les agriculteurs sont absents de la cohorte, et la part des ouvriers et des employés diminue au profit des professions intermédiaires et libérales.

Tableau 5.5. Panel SN00-02 : CSP des participants de SN00-02

Occupation	SN00-02		Internautes français, déc. 2001
	En 2000	En 2002	
Retraité	3,8 %	4,6 %	6,2 %
Sans profession, au foyer	11,4 %	5,8 %	9,4 %
Temporairement sans emploi (chômage, maladie...)	2,5 %	3,1 %	1,6 %
Étudiant	23,4 %	23,3 %	23,3 %
Ouvrier	1,8 %	1,7 %	4,4 %
Agriculteur, pêcheur	0 %	0 %	0,3 %
Employé	15,0 %	16,5 %	18,8 %
Profession intermédiaire (cadre, chef de service, technicien, etc.)	31,5 %	33,7 %	24,2 %
Profession libérale	1,5 %	2,1 %	1,5 %
Dirigeant (PDG, DG, directeur, cadre supérieur...)	6,9 %	7,1 %	8,4 %
Propriétaire d'une entreprise, commerçant	2,1 %	2,1 %	1,5 %

Pour ce qui est de l'analyse des parcours Internet, attendons-nous également à ce que les contenus visités reflètent en partie les centres d'intérêt d'une population au pouvoir d'achat moyen ou élevé, et que certains services et modes de communication soient sous-représentés.

Ces données présentent un intérêt très important : il n'existe pas à ce jour d'étude d'usages d'Internet portant sur une période d'observation aussi longue, ni sur une cohorte aussi importante. De telles données permettent de répondre à des questions jusqu'alors en suspens, en particulier sur l'évolution des pratiques au fil du temps et, pour le cadre plus précis de la navigation Web, l'établissement de routines et de visites fréquentes de certains sites.

Synthèse. *Le panel SensNet 2000-2002 offre une durée d'observation inédite d'une cohorte d'internaute résidentiels. De telles données longitudinales autorisent notamment une analyse de la structure et de l'évolution des territoires personnels sur le Web.*

5.1.3 Panel BibUsages

Le dernier lot de données de trafic est apporté par le projet BibUsages, mené par France Télécom R&D et la Bibliothèque Nationale de France, qui vise à étudier l'usage des bibliothèques électroniques en ligne par le grand public[1]. Si ces données ne sont pas représentatives des pratiques Web en général, leur intérêt est ailleurs : il s'agit de donner un éclairage particulier sur des pratiques non observables dans les échantillons précédents du fait du peu d'internautes concernés. En outre, on perd en quantité ce que l'on gagne en qualité : la méthodologie menée ici est plus complète que pour les panels SN2002 et SN00-02, et permet de mener des études qualitatives fines.

Phases du projet et méthodologie

Le projet BibUsages s'est déroulé en trois phases :

1. Enquête de cadrage *via* un questionnaire en ligne (mars 2002).
2. Constitution d'un panel et recueil de trafic Web pour ce panel (avril-décembre 2002).
3. Enquête qualitative par entretiens (octobre 2002).

Dans la première phase de l'expérimentation, un questionnaire a été soumis aux visiteurs du site Gallica en mars 2002 durant trois semaines. Il a permis à la fois d'avoir une connaissance plus précise du public de Gallica, et de recruter les volontaires pour faire partie du panel d'utilisateurs dont le trafic Web a été enregistré. Le questionnaire utilisé est fourni en annexe (voir Annexe 4).

Outre les caractéristiques socio-démographiques des répondants, le questionnaire s'articule autour de deux thématiques principales : d'une part, l'usage de Gallica (fréquence des visites, rubriques consultées, etc.), et d'autre part les usages d'Internet en général (intensité d'usage, services utilisés, types de sites visités, etc.). À la fin du questionnaire, les répondants se sont vus proposer de participer au panel d'utilisateurs mis en place. Au terme de cette première étape, 2 340 personnes ont répondu au questionnaire, et 589 ont donné leur accord de principe pour faire partie du panel d'utilisateurs, soit près d'un quart.

Dans un deuxième temps, un panel représentatif de la population totale des répondants au questionnaire a été constitué, composé de 72 volontaires qui ont téléchargé et installé le dispositif de recueil de trafic. Les premières données de trafic nous parviennent fin mai 2002, mais l'installation du dispositif de recueil n'est achevée sur la quasi-totalité des postes que début juillet. À partir de juillet, l'ensemble des volontaires avaient installé le dispositif de recueil de trafic sur leur

[1] Projet soutenu par le Réseau National de Recherche en Télécommunications ; voir Annexe 1 pour une description complète du projet.

poste ; leur activité Web a été enregistrée pendant six mois, jusqu'en décembre 2002[1].

En complément des données de trafic, des entretiens semi-directifs ont été menés auprès de 16 des 72 participants à l'expérimentation[2]. Les entretiens ont été organisés en juillet et réalisés en octobre 2002. Sur les 72 membres actifs du panel, 50 ont été contactés afin d'obtenir leur accord de participation aux entretiens, et ce en s'intéressant aux membres les plus actifs en fonction de leur trafic Internet.

Le questionnaire de Gallica de mars 2002 et le trafic enregistré par Audinet[3] permettent de dégager des profils d'usages pour chaque utilisateur interviewé, en particulier autour des types de contenus visités et de l'intensité des pratiques. Néanmoins les modalités et les contextes d'utilisation restent à approfondir. Les entretiens comblent ce vide en permettant d'une part de confirmer les pratiques telles qu'elles émergent de l'analyse du trafic et d'autre part de les inscrire dans leur contexte (usages d'Internet en général – et pas seulement du Web – et les pratiques hors-ligne). Il s'agit ainsi de voir comment la consultation des bibliothèques numériques s'inscrit dans une pratique générale d'Internet dans un contexte donné, et de mieux connaître les différents types d'usages, les motivations et les modalités de la pratique avec comme appui les données de trafic.

Pour mener ces entretiens, une grille d'entretien reprenant en filigrane ces objectifs, fournie en annexe (voir Annexe 4, Matériau d'enquête BibUsages), a été élaborée autour de trois axes :

1. Une première partie centrée sur l'utilisation générale d'Internet permet de mieux cibler le profil général de l'internaute dans ses pratiques : durée, motivations, contexte de l'utilisation, modalités des recherches et traitement de l'information. En outre, cela permet d'avoir des renseignements sur les usages hors Web (*chat*, mail, forums, *peer-to-peer*…) que la sonde Audinet, dans la version utilisée pour cette étude, n'enregistre pas.
2. La deuxième partie de la grille se concentre sur l'usage des bibliothèques électroniques et de Gallica en particulier. Il s'agit de connaître les contextes d'utilisation des fonds numériques, les méthodes de recherche et les modalités de traitement de l'information. Dans la discussion, on cherche également à pressentir des difficultés et à obtenir des propositions d'amélioration dans la conception du site Gallica.
3. La troisième partie de l'entretien se concentre sur les pratiques « off-line » : il s'agit ici de relier l'utilisation des bibliothèques électroniques à celle des

---

[1] Nous tenons particulièrement à remercier les personnes qui ont accepté de participer à cette étude, en installant le dispositif de recueil de trafic et en participant aux entretiens que nous avons menés. C'est grâce à leur collaboration et au temps qu'ils ont consacré à ce projet que celui-ci a pu être mené à bien.

[2] La plupart des entretiens ont été menés par France de Charentenay (BnF) ; nous avons mené ou participé à la plupart d'entre eux.

[3] Sonde de recueil de trafic Internet développé par France Télécom R&D ; voir chapitre 2.1.1, « Technologies de recueil de données ».

bibliothèques classiques et, plus largement, aux pratiques de lecture et aux pratiques culturelles des interviewés.

Pour chaque entretien, une fiche descriptive du panéliste a été élaborée à partir de ses réponses au questionnaire de Gallica (mars 2002) et des statistiques de son trafic Internet déjà recueilli *via* Audinet. En s'appuyant sur ces informations, l'entretien permet de confirmer et d'expliciter les pratiques observées.

La richesse de la méthodologie suivie réside dans l'exploitation conjointe et croisée des données issues des trois phases du projet. Le questionnaire donne une description précise de la population d'ensemble et nous permet ainsi de bien situer l'analyse plus fine des phases 2 et 3 dans un cadre général. L'approche qualitative permet de valider des hypothèses issues de la phase d'analyse de trafic, de même que l'analyse de trafic permet de consolider les conclusions issues de l'analyse des entretiens, en les appuyant sur des mesures objectives des usages.

Profil général à l'issue du questionnaire

En premier lieu, on note une part importante de visiteurs étrangers : parmi les répondants à l'enquête, 31,4% déclarent résider à l'étranger ; en outre, pour 60% d'entre eux, le français n'est pas leur langue maternelle (mais ils la pratiquent assez pour répondre au questionnaire). Gallica s'impose ainsi comme un point d'accès à des fonds francophones depuis l'étranger.

Dans le « lectorat » de Gallica, tel qu'il ressort de l'enquête, on constate les tendances fortes suivantes :

- un niveau d'études élevé : 33,8 % entre Bac+2 et Bac+4, et 38,3 % de diplômés de troisième cycle ;
- une représentation majoritaire des plus de cinquante ans, qui comptent pour 32,7 % des répondants, au détriment des moins de trente ans (11,8 %) ;
- une sur-représentation des cadres de la fonction publique, très loin devant les autres catégories (l'enseignement supérieur est le secteur d'activité le plus représenté).

En ce qui concerne les usages d'Internet, les personnes interrogées déclarent un haut degré de pratique, utilisant les services Web de façon très régulière (recherche d'informations ou opérations bancaires ou boursières, achats en ligne), tout cela majoritairement à partir d'un accès à domicile. En même temps, il faut remarquer que ces mêmes utilisateurs sont d'anciens internautes (37% d'entre eux sont connectés à Internet depuis 1997 ou antérieurement). On voit donc que le profil général des utilisateurs rencontrés dans cette enquête se situe dans un contexte de régularité et de fidélité, de connaissance de l'outil, avec un usage personnel à partir du domicile et dans des sessions plutôt longues.

Pour analyser la spécificité du public de Gallica par rapport aux internautes français, nous recourons aux chiffres données par NetValue en décembre 2001, que nous comparons aux caractéristiques des répondants au questionnaire résidant en France.

Tableau 5.6. « Gallicanautes » résidant en France et internautes français

		France - données NetValue décembre 2001	Répondants BibUsages résidant en France (mars 2002)
Homme / femme		58% / 42%	69,3% / 30,7%
Urbain		81,0%	86,4%
Âge	Moins de 15 ans	4,6 %	0,5 %
	15-24 ans	24,7 %	9,3 %
	25-34 ans	23,1 %	19,2 %
	35-49 ans	31,8 %	32,8 %
	50-64 ans	12,5 %	31,1 %
	Plus de 65 ans	3,3 %	7,1 %
Bas débit / haut débit		91,1% / 8,9%	67,2% / 32,8%

Les caractéristiques générales (voir Tableau 5.6) montrent que les gallicanautes sont, par rapport aux internautes français, plutôt des hommes, globalement plus âgés (sur-représentation des tranches 50-64 ans et plus de 65 ans). L'écart le plus important concerne le type d'équipement Internet : le haut débit est fortement sur-représenté, avec un taux d'équipement de près de 33% chez les utilisateurs de Gallica, contre 9% en général à la même époque. Les gallicanautes sont également des internautes plus « anciens », avec 67,2% des répondants résidant en France disposant d'une connexion depuis 1999 au moins.

Tableau 5.7. Spécificités des gallicanautes résidant en France : CSP[1]

France - données NetValue (déc. 2001)		Questionnaire BibUsages en ligne (mars 2002)	
Retraité	6,2 %	Retraité	10,1 %
Sans profession, au foyer	9,4 %	Au foyer	1,3 %
Temporairement sans emploi	1,6 %	Temporairement sans emploi	2,5 %
Étudiant	23,3 %	Étudiant	18,4 %
Ouvrier	4,4 %	Ouvrier	0,3 %
Agriculteur, pêcheur	0,3 %	Agriculteur, exploitant	0 %
Employé	18,8 %	Employé, personnel de service	6,5 %
Profession intermédiaire	24,2 %	Technicien, agent de maîtrise, cat. B	10,0 %
Profession libérale	1,5 %	Profession libérale	6,0 %
Propriétaire d'une entreprise	1,5 %	Commerçant, artisan	0,5 %
		Cadre du secteur privé	13,8 %
Dirigeant	8,4 %	Cadre de la fonction publique (cat. A)	26,5 %
		Chef d'entreprise, cadre dirigeant	4,0 %

L'examen des professions et catégories socio-professionnelles des répondants au questionnaire résidant en France (voir Tableau 5.7) complète l'analyse en montrant une sur-représentation des CSP supérieures (cadres, enseignants), tandis qu'au sein des étudiants, les 2e et 3e cycle sont sur-représentés parmi les visiteurs de Gallica.

[1] Les grilles de CSP employées par NetValue et dans le questionnaire BibUsages étant différentes, nous proposons des équivalences, sans garantir leur concordance.

Composition du panel BibUsages

Le panel BibUsages, composé d'internautes chez qui nous avons installé le dispositif de recueil de trafic Web, a été composé de telle manière qu'il reflète le plus fidèlement possible les caractéristiques des répondants au questionnaire présenté sur Gallica en mars 2002. Cependant, en raison de contraintes d'organisation, le panel est exclusivement composé de personnes résidant en France : les répondants étrangers à l'enquête n'y sont donc pas du tout représentés. Nous donnons ici les caractéristiques de ce panel, qui reprend de manière générale celles des « gallicanautes » présentées précédemment.

En termes socio-démographiques, le panel BibUsages est essentiellement masculin (¾ d'hommes). La moyenne d'âge est de 46 ans ; toutefois, cette moyenne ne met pas en évidence la sur-représentation des panélistes de plus de 55 ans, avec plus de 31 % du panel se situant dans les deux dernières tranches d'âge contre 19 % pour l'enquête de Gallica. Les panélistes exercent majoritairement une activité professionnelle (82 %) et pour 45 % d'entre eux, occupent un poste cadre de la fonction publique, dans le secteur de l'enseignement ou de la recherche. Dans le cadre des panélistes n'ayant pas d'activité rémunérée (18 %), plus de 60 % sont des retraités, alors qu'ils n'étaient que 23,4 % dans l'enquête générale. De plus il faut noter que 20 % d'entre eux sont des étudiants de deuxième cycle. En ce qui concerne la provenance géographique du panel, la diversité des régions françaises a été respectée avec une sur-représentation de Paris et la région parisienne.

Plus de 47 % des panélistes utilisent Internet depuis 1997, de manière quotidienne et dans la majorité des cas de leur domicile (plus de 62 %). Ils sont largement équipés d'accès haut-débit, puisque 29 des 72 participants ont une connexion par ADSL ou Câble. Le questionnaire en ligne nous apprend que les panélistes utilisent principalement Internet pour rechercher des informations (30 %) et échanger à partir des différents modes de communication (mail, *chat*, forum). Leurs principaux centres d'intérêts sur le Web sont assez hétérogènes, avec une prédominance de l'art et de la littérature, des sciences sociales et de la recherche documentaire ou bibliographique.

Le panel est une population d'internautes fidèles à Gallica : plus de 47 % déclarent en être à plus de dix visites. Ils fréquentent le site de manière régulière et pour la plupart de chez eux. Leurs visites sont d'une durée (déclarée) supérieure à 10 minutes : 38 % disent rester entre 10 à 30 minutes et 32 % plus d'une demi-heure. Les panélistes ont connu Gallica par différents moyens : brochure de la Bibliothèque Nationale de France, lien d'un autre site, requête *via* un moteur de recherche ou annuaire Web.

Profil général des personnes interrogées en entretien

Le profil socio-démographique des seize personnes interrogées reste dans la lignée du panel, ainsi que de la population générale de l'enquête de Gallica (voir Tableau 5.8 ci-dessous). La plupart des interviewés occupent des postes de cadre de la fonction publique ou du secteur privé ; la moyenne d'âge se situe autour de 48 ans avec une fourchette allant de 33 à 76 ans, et la majorité habite en milieu urbain.

Tableau 5.8. Récapitulatif des profils des panélistes interrogés en entretien

Utilisateur	Sexe	Département	Tranche d'âge	Profession	Type de connexion à domicile	Ordinateur partagé avec d'autres membres de la famille	Nbre total de sessions réalisées	Lieu d'installation d'Audinet Domicile/ Travail	Ancienneté Internet
Utilisateur A	F	16	55/59	Enseignante du supérieur	Haut débit	Oui	905	D	1998
Utilisateur B	F	75	30/34	Indépendante Correctrice	Haut débit	Oui	110	D/T	1997
Utilisateur C	F	35	55/59	Cadre de la fonction publique	Haut débit	Non	pb. install.	D	1997
Utilisateur D	H	94	35/39	En Formation Sciences Sociales	Haut débit	Oui	642	D	1997
Utilisateur E	H	83	70/74	Retraité (cadre du privé)	RTC	Oui	61	D	2000
Utilisateur F	H	77	50/54	Enseignant du primaire	RTC	Oui	150	D	1998
Utilisateur G	H	75	40/44	Enseignant (Histoire)	Haut débit	Oui	626	D	1999
Utilisateur H	H	75	45/49	Cadre du privé (informatique)	Haut débit	Oui	28	D	1999
Utilisateur I	H	83	45/49	Employé, personnel de service	RTC	Oui	667	T	1998
Utilisateur J	H	33	50/54	Cadre de la fonction publique	Numéris	Non	278	D	2000
Utilisateur K	H	27	30/34	Cadre du privé	Haut débit	Non	36	T	1997
Utilisateur L	H	69	30/34	Cadre de la fonction publique (CNRS)	RTC	Non	218	T	1997
Utilisateur M	H	59	50/54	Enseignant du supérieur	Haut débit	Oui	381	D	1997
Utilisateur N	F	92	55/59	cadre du privé (assurance)	RTC	Non	262	D	1997
Utilisateur O	H	91	75/79	Retraité (comptable)	RTC	Non	658	D	2000
Utilisateur P	H	15	65/69	Retraité (enseignant du secondaire)	Haut débit	Non	101	D	1997

En ce qui concerne l'utilisation d'Internet, on remarque qu'il s'agit d'anciens internautes, avec en moyenne déjà plus de 3 ans de pratiques, et ayant un usage fréquent et une consommation importante (déclarée dans le questionnaire en ligne, et confirmée par les données de trafic).

Les participants ont en général installé Audinet chez eux. Pour certains, cet ordinateur est un ordinateur familial dont ils sont l'utilisateur principal ; il ressort des entretiens que l'utilisation par les autres membres du foyer reste occasionnelle. Les choix individuels se sont portés sur l'équipement le plus utilisé pour l'installation de la sonde, c'est pourquoi seules deux personnes ont installé ce logiciel sur leur lieu de travail.

C'est sur les données de trafic issues de ces trois panels – SN2002, SN00-02 et BibUsages – que nous appliquons nos outils et nos méthodes d'enrichissement et d'analyse de trafic afin de décrire les parcours des utilisateurs sur le Web. Toutefois, nous ne mettons pas ces trois sources de données sur le même plan : le panel SN2002 est le point focal de notre analyse, car il s'agit à la fois du plus récent, du plus volumineux et du plus représentatif des trois. Les panels SN00-02 et BibUsages sont mobilisés en appui, pour vérifier si les résultats obtenus sur le panel de référence sont observables chez ces deux autres populations. En outre, les données sur trois ans sont mobilisées pour les études longitudinales, afin d'observer des effets d'habitude et de constitution de territoires, tandis que les données BibUsages nous permettent de faire un éclairage ciblé sur une population atypique d'internautes anciens, intensifs, fortement équipés en haut débit et consommateurs de contenus culturels.

> ***Synthèse.*** *Le panel BibUsages, de plus petite taille que les précédents, est centré sur la population des internautes visiteurs de bibliothèques numériques en ligne. Sa construction suit une méthodologie en « entonnoir » : un questionnaire cerne le profil général de cette population, dont un panel est extrait pour faire du recueil de trafic, et des entretiens avec seize panélistes viennent compléter qualitativement ces données. Les trois panels (SN2002, SN00-02 et BibUsages) sont complémentaires, et chacun répond à des problématiques spécifiques.*

5.1.4 Usages généraux d'Internet

Avant d'entrer plus précisément dans l'analyse des sessions Web et des parcours de site en site, nous souhaitons nous doter de descriptions sur les usages généraux d'Internet au sein de nos trois panels. Ces données doivent servir de cadrage pour ce qui va suivre, et replacent les parcours Web dans la perspective plus générale de l'accès au réseau. Car si nous avons fait le choix ici de ne traiter que des parcours sur le Web, nous n'oublions pas qu'une approche plus vaste se doit d'analyser l'ensemble de l'activité de chaque utilisateur en considérant que les ruptures techniques entre les dispositifs qui sous-tendent l'Internet ne sont pas forcément perçues par l'utilisateur et qu'il y a pour lui une véritable continuité entre Web, messagerie, forum, etc. autour d'un objet unique, l'ordinateur.

Panels SensNet : profils d'usage différenciés

Les données SensNet concernent l'ensemble des protocoles mobilisés dans l'accès à Internet : on dispose d'informations sur l'activité Web, mais aussi sur la messagerie, les jeux en ligne, le *chat*, l'échange de fichiers, etc. Ces traces complètes de l'activité sur le Net permettent de construire une description des internautes en fonction de leur utilisation ou non de ces différents services, et de l'intensité d'usage qui y est attachée.

Pour cela, le panel SensNet 2002, représentatif des internautes sur les dix mois d'observation, nous sert de trame de fond : pour chaque panéliste, on examine le nombre de sessions pour chaque type d'activité (Web, mail, *peer-to-peer*, etc.) de janvier à octobre 2002. Le travail de classification sur cette base est compliqué par le fait que les variables ne sont pas homogènes : certains types d'activité comme le Web ou la messagerie sont très présents et partagés par la quasi-totalité des internautes, tandis que le *chat* ou le téléchargement concernent peu d'individus, et impliquent moins de sessions. Pour contourner cette difficulté, nous avons discrétisé chaque variable de manière *ad hoc*, en tenant compte à la fois du fait d'avoir utilisé ou non un protocole, et du nombre de sessions que son utilisation implique à l'échelle du panel (voir Tableau 5.9). Certains services comme les forums sont ainsi discrétisés en « utilisation / pas d'utilisation », tandis que l'usage du Web est décomposé en trois modalités : peu intensif en deçà de 20 sessions sur les dix mois, intensif au-delà de 130 sessions.

Tableau 5.9. Variables et modalités pour la description des usages d'Internet

Variable	Modalités	Intervalles (nb de sessions) sur 10 mois	Répartition des panélistes
Chat/IRC	Pas de chat/IRC	0	70,4 %
	Chat/IRC faible	1-2	13,0 %
	Chat/IRC intense	3 et plus	16,6 %
Web	Web - peu	0-19	29,2 %
	Web - moyen	20-129	41,4 %
	Web - intense	130 et plus	29,4 %
WebMail	Pas de WebMail	0	28,1 %
	WebMail - peu	1-4	24,6 %
	WebMail - moyen	5-39	29,6 %
	WebMail - intense	40 et plus	17,7 %
WebChat	Pas de WebChat	0	67,4 %
	WebChat - faible	1-2	15,3 %
	WebChat - intense	3 et plus	17,3 %
Mail	Pas de mail	0	31,5 %
	Mail - peu	1-8	28,4 %
	Mail - moyen	9-69	25,8 %
	Mail - intense	70 et plus	14,2 %
Peer to peer	Pas de p2p	0	78,1 %
	p2p faible	1-15	12,8 %
	p2p intense	16 et plus	9,2 %
Forums	Pas de forum	0	92,8 %
	Forum	1 et plus	7,2 %
Téléchargement	Pas de téléchargement	0	57,8 %
	Téléchargement faible	1-2	20,0 %

	Téléchargement moyen	3-8	13,3 %
	Téléchargement intense	9 et plus	8,9 %
Audio-vidéo	Pas d'audio-vidéo	0	53,0 %
	Audio-vidéo faible	1-2	20,2 %
	Audio-vidéo moyen	3-8	13,2 %
	Audio-vidéo intense	9 et plus	13,6 %
Jeux	Pas de jeux	0	90,6 %
	Jeux	1 et plus	9,4 %

Clef de lecture : un panéliste ayant réalisé deux sessions de Chat/IRC sur les dix mois d'observation sera rattaché à la modalités des « faibles utilisateurs » de ce service.

Chaque panéliste est ainsi décrit par 10 variables discrètes relatives à son intensité d'usage de chaque type de protocole sur Internet. L'analyse en composantes multiples pratiquée sur cette base et la classification sur les dix facteurs premiers nous permet d'identifier quatre classes d'utilisateurs distinctes, que représente la Figure 5.1 ci-dessous[1].

Un premier groupe d'utilisateur se distingue par la faible intensité et le peu de diversité de ses usages (classe n°2 sur le graphique). Chez ces utilisateurs occasionnels, qui représentent 33 % du panel, on ne compte que 2,9 sessions par mois en moyenne (médiane : 1 session), contre 145 pour l'ensemble du panel (médiane : 70). Les services utilisés se limitent alors aux « fondamentaux » : leur activité sur le Net se résume à un peu de Web, et un peu de messagerie.

Un second groupe constitué de 35 % du panel (classe n°1 sur le graphique) présente des usages à la fois plus intenses et plus diversifiés. Cette classe des « internautes ordinaires », qui compte 9,3 sessions par mois en moyenne (médiane : 7,2) a une activité moyenne en ce qui concerne le Web, le mail et le WebMail ; un peu de téléchargement et de contenus audio/vidéo complètent ce menu où l'on ne trouvera ni *peer-to-peer*, ni outils avancés de communication (*chat*, WebChat, forum).

Ces services sont l'apanage d'un troisième groupe d'internautes (classe n°3, 22 % du panel), chez qui la pratique des outils de communication interpersonnelle est particulièrement intense. Corrélativement, l'usage du Web est également très fort chez ces « internautes communicants », qui pratiquent autant la messagerie et le *chat* sur le Web que sur les outils dédiés.

Le dernier groupe, comptant pour 10 % du panel, s'oriente bien plus vers des activités ludiques : le *peer-to-peer*, absent ou très faible dans les trois autres groupes, est ici particulièrement intense, de même que les services audio/vidéo et le téléchargement de fichiers. En ce qui concerne la messagerie instantanée, ces internautes intensifs se distinguent non seulement par une utilisation forte, mais surtout le délaissement du WebChat au profit du *chat* sur logiciel spécifique (ICQ, Messenger, etc.). Sans être déterminante, la pratique des jeux en réseau est également forte dans ce groupe, où elle touche 40 % des individus, contre 10 % dans l'ensemble du panel. Ces utilisateurs intensifs sont enfin ceux qui utilisent le plus le Web, avec 39 sessions par mois en moyenne pour chaque panéliste.

[1] Nous utilisons pour ces calculs le logiciel SPAD (www.cisia.com).

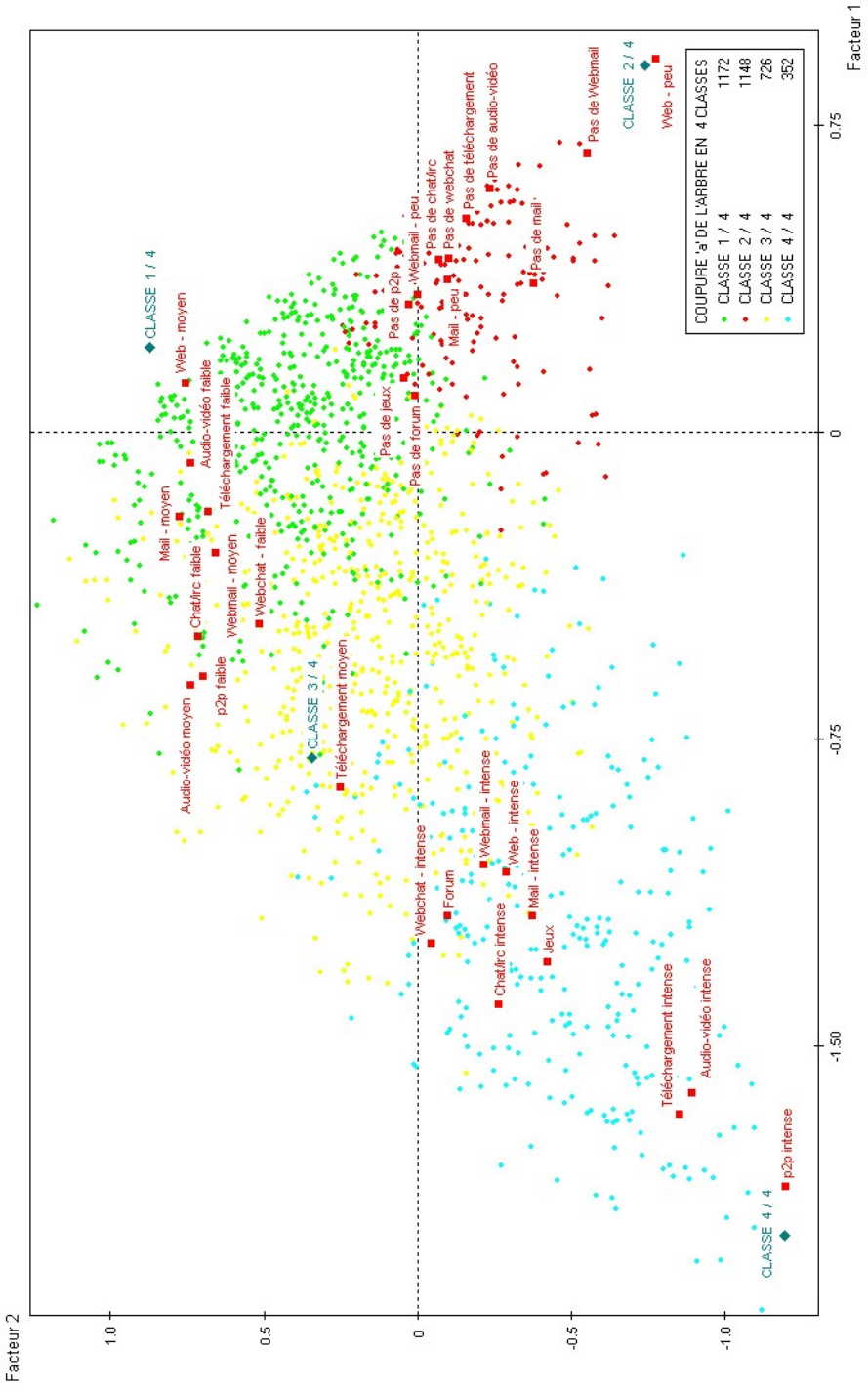


Figure 5.1. Classification des panélistes SN2002 sur la base de leur activité Internet - 4 classes, axes 1 et 2

Les internautes du panel SN00-02 suivis sur trois années constituent, en 2002, un sous-groupe du panel représentatif SensNet 2002. On observera avec intérêt dans quel groupe-type d'usagers ces internautes anciens se positionnent en 2002 : on peut supposer que plus de trois années d'utilisation suivie d'Internet impliquent un ancrage et une intensification des pratiques. Il ne semble pourtant pas que ce soit toujours le cas : la répartition des internautes du panel SN00-02 dans les quatre catégories-types d'utilisateurs suit globalement celle du panel général SN2002 (voir Tableau 5.10).

Tableau 5.10. Répartition des individus dans les groupes-types d'internautes

	Panel SN2002	Sous-panel SN00-02
Utilisateurs occasionnels	33,3 %	24,2 %
Internautes ordinaires	34,8 %	34,8 %
Communication interpersonnelle	21,5 %	27,7 %
Usages ludiques	10,4 %	13,3 %
Total	100 %	100 %

Ces internautes « anciens » sont plus actifs que les autres : ils sont globalement plus présents dans les trois groupes d' « internautes ordinaires », « communication interpersonnelle » et « usages ludiques » que l'ensemble du panel SN2002, au détriment des usages occasionnels. Cependant, un quart d'entre eux se rattache à cette dernière classe, contre un tiers dans l'ensemble du panel SN2002. Ce résultat est intéressant : il montre que l'ancienneté n'est pas un moteur systématique d'intensité d'usage, et que l'utilisation d'Internet peut être ancrée dans les pratiques des individus tout en demeurant une ressource mobilisée de façon occasionnelle.

BibUsages : utilisateurs avertis pour usages ciblés

Le dispositif de recueil de trafic utilisé dans BibUsages a été paramétré pour ne recueillir que les données relatives à l'activité Web, en conséquence de quoi il nous est impossible de pratiquer sur ce panel la même segmentation que pour les données SensNet. Cela étant, nous disposons des 2 340 réponses au questionnaire soumis en ligne, qui permettent de dresser un panorama rapide des usages d'Internet des gallicanautes.

En premier lieu, les répondants au questionnaire se présentent comme des internautes très réguliers et très fidèles : 73 % d'entre eux déclarent utiliser Internet quotidiennement, tandis qu'ils ne sont que 1,5 % à moins de 3 connexions par mois (voir Tableau 5.11).

Tableau 5.11. Questionnaire BibUsages : fréquence d'utilisation d'Internet

À quelle fréquence utilisez-vous Internet ?	% du total
Tous les jours	72,9 %
2 à 5 fois par semaine	21,4 %
Environ une fois par semaine	4,2 %
1 à 3 fois par mois	1,2 %
Moins souvent	0,3 %

En ce qui concerne les usages des différents services disponibles sur Internet (voir Tableau 5.12), les visiteurs de Gallica interrogés semblent se placer plutôt dans

le groupe des « internautes ordinaires » : la recherche d'information sur le Web est pratiquée par la quasi-totalité des répondants, et certaines pratiques comme l'achat ou le recours aux services bancaires en ligne dénotent des usages avancés du Web.

Tableau 5.12. Questionnaire BibUsages : usages d'Internet

Quel usage avez-vous d'Internet ? (plusieurs réponses possibles)	% du total
Recherche d'information	98,8 %
Communication : *chat*, groupes de discussion, messagerie, ...	57,9 %
Achat ou opération financière, dont	48,2 %
- Achat en ligne	34,2 %
- Opérations et consultations bancaires ou boursières	33,1 %
Téléchargement, dont	40,1 %
- Téléchargement de logiciels	32,6 %
- Téléchargement de musique et/ou de vidéo	21,4 %
Jeux en ligne	5,7 %
Autre usage	9,3 %

L'absence d'informations sur l'intensité d'usages des différents services nous empêche d'avoir une vue plus précise : la pratique du téléchargement, qui touche 40 % des répondants, dont la moitié pour des fichiers audio/vidéo, pourrait rattacher les gallicanautes au groupe des « internautes ludiques », mais la faible part du jeu dans les pratiques laisse penser que ce n'est pas le cas, et l'on peut supposer que ces pratiques de téléchargement restent marginales en termes d'intensité.

Enfin, en ce qui concerne les outils de communication, ceux-ci sont utilisés par 58 % des répondants, mais la généricité de cette catégorie ne permet pas de conclure à la place de la communication dans les pratiques des répondants. Les entretiens avec seize membres du panel BibUsages laissent penser que celle-ci reste secondaire : au fil des discussions, il est apparu que l'usage de la messagerie n'arrive pas dans les premières considérations des interviewés. Elle est certes importante mais ne rentre pas dans les motivations personnelles d'accès à Internet : elle permet d'échanger essentiellement avec la famille géographiquement éloignée et les spécialistes rencontrés dans le cadre des recherches. Ce volet orienté vers la communication interpersonnelle est complété par un usage relatif à la recherche d'information, *via* l'abonnement pour certains à des *newsletters* et à des listes de diffusion[1].

Pour les trois panels, l'analyse générale des usages d'Internet montre des variations fortes en termes d'intensité entre les différents individus ; elle met surtout en avant le fait que le Web, à la différence d'autres services comme le *peer-to-peer* ou le *chat*, s'impose comme l'outil fondamental et le pivot de l'activité sur Internet. Non seulement le Web est le seul service mobilisé par les faibles utilisateurs, mais il accompagne également le recours à des outils plus spécifiques (audio/vidéo, téléchargement, etc.) où il est utilisé de manière intensive. L'analyse des parcours

[1] Nous n'avons pas particulièrement mis l'accent sur les usages des outils de communication dans cette étude : le dispositif de recueil de trafic utilisé n'était pas paramétré pour enregistrer cet usage ; par ailleurs, nous n'avons pas cherché à développer cet aspect lors des entretiens.

Web doit tenir compte de cette diversité, qui implique que la navigation s'insère dans des pratiques et des contextes d'usage bien différenciés.

> ***Synthèse.*** *La segmentation des internautes sur la base de l'intensité d'usage des différents outils Internet (Web, mail, messagerie instantanée, jeux, etc.) fait émerger quatre profils distincts : les utilisateurs occasionnels, qui font uniquement du Web et du mail en petite quantité ; les internautes « ordinaires », plus intensifs mais peu diversifiés dans leurs usages ; les communicants, qui ont régulièrement recours aux outils de communication interpersonnelle (*chat, *messagerie instantanée) ; et les utilisateurs ludiques, plutôt orientés vers les jeux ou le* peer-to-peer. *Mobilisé par les quatre groupes, le Web s'impose comme outil pivot de l'activité sur Internet, mais il s'insère, selon les pratiques des individus, dans des contextes d'usages différenciés.*

5.2 Volumétrie, temporalité et topologie des parcours

En première approche des parcours sur le Web, on s'intéressera aux éléments temporels et topologiques de la navigation : ceci permet d'appréhender la diversité des parcours, et de les replacer dans le contexte d'une description de l'activité de navigation préalable à l'examen des contenus et des services visités.

5.2.1 Intensités d'usage variées

Nous présentons ici les résultats de l'analyse des données de trafic en termes de nombre de pages, de sites, et de durée. L'objectif est de donner, avant d'aller plus loin, un panorama général du trafic réalisé par les membres de nos trois panels. En arrière-plan, il s'agit de répondre à une question d'ordre méthodologique : quelles informations peut-on extraire en termes d'usages de données non qualifiées sur le plan du contenu ? Les URL visitées sont ici manipulées comme des données symboliques, regroupées en sites, sans indication de contenu.

Volumétrie globale du trafic des différents panels

De manière globale, nos trois sources de données représentent près de 45 années de navigation Web cumulée répartie en plus de 681 000 sessions. Ces éléments de volumétrie cumulée, quelque peu absurdes en eux-mêmes, ne témoignent que de la taille importante des données à traiter. Ils montrent ici la nécessité de se doter d'outils d'analyse permettant de manipuler des données en masse, à travers l'élaboration d'agrégats et d'indicateurs, tout autant que d'aller dans le détail de chaque session.

Tableau 5.13. Trafic des trois panels : volumétrie générale

	BibUsages	SN00-02	SN2002
Nombre d'utilisateurs	72	597	3372
Nombre de sessions	17 083	261 634	403 129
Durée d'observation	6 mois	34 mois	10 mois
Nb moyen de sessions par mois et par util.	39,5	12,7	11,9
Nb pages vues (en millions)	1,41	15,66	26,72
Nb pages distinctes (en millions)	0,65	4,60	6,75

Les volumes globaux par jeu de donnée montrent des disparités importantes (voir Tableau 5.13) : pour le plus important, SN2002, près de 27 millions de pages sont vues au cours de 403 000 sessions Web, contre 1,4 millions en 17 000 sessions pour BibUsages. Ces éléments ne doivent pas masquer les véritables disparités entre échantillons en termes d'usages : si le trafic global de BibUsages est bien moindre que celui des panels SensNet, du fait du nombre de participants et de la durée d'observation, les utilisateurs de BibUsages sont plus intensifs que les autres. Ainsi, chaque panéliste de BibUsages réalise en moyenne 39,5 sessions par mois (médiane : 23,2), tandis que ceux de SN2002 en font en moyenne 11,9 (médiane : 5,7). Les utilisateurs du panel 2000-2002, dont nous avons vu qu'ils étaient majoritairement des « anciens internautes », sont plus proches des seconds que des premiers, avec 12,7 sessions par mois en moyenne pour chacun (médiane : 8,7).

À l'examen de ce premier résultat, il semblerait bien que l'ancienneté ne soit pas un facteur déterminant dans l'intensité de la pratique, mais que les centres d'intérêt – en l'occurrence, l'attrait pour les bibliothèques électroniques – soit un moteur d'intensité d'utilisation. Ceci est sûrement en partie vrai, mais doit être tempéré par un autre élément : nos trois panels sont résidentiels, mais dans le cas de BibUsages, nous avons moins d'actifs, sinon beaucoup d'actifs exerçant une partie de leur activité à domicile.

Un examen longitudinal permet de donner plus de profondeur à l'analyse de l'activité des panels : la pratique du Web tend-elle à s'intensifier au fil du temps, ou se stabilise-t-elle ? Deux hypothèses opposées sont possibles : d'une part, on peut supposer que l'accroissement et la diversification de l'offre de contenus et de services sur le Web amènent les internautes à intensifier leurs pratiques ; d'un autre côté, on peut postuler l'ancrage des pratiques autour de certains sites amenant une stabilisation globale des usages au fil du temps. La répartition de l'activité des trois panels tout au long des périodes d'observation montre des comportements assez différenciés, qui tiennent plus à la composition et au maintien des panels. Dans le cas de BibUsages, on constate, malgré une activité relativement soutenue de juin à décembre, une baisse constante du nombre de panélistes actifs par mois, qui passe de 65 au début de la période à 40 en décembre 2002. Ceci est principalement dû au fait que certains participants à l'expérimentation ont désinstallé la sonde Audinet et quitté le panel sans que nous en soyons informés.

Une baisse similaire de l'activité, quoique moindre, est observable pour les données issues du panel SensNet 2002 : le nombre de panélistes actifs par mois décroît doucement (2 500 en janvier, 2 200 en octobre), de même que le nombre de sessions par mois. Cependant, les utilisateurs quittant le panel ont été filtrés dans ces

données, et la baisse d'activité est ici imputable aux « abandonnistes », des utilisateurs dont l'usage d'Internet s'affaiblit jusqu'à cesser. Dans le cadre du projet TypWeb[1], portant sur 1 140 internautes observés tout au long de l'année 2000, ce phénomène avait déjà été observé et décrit : « Pour une fraction des internautes, les usages sont rares et tendent à se raréfier, voire à disparaître au fil des mois ; pour les autres au contraire, surtout pour les gros utilisateurs, la pratique tend à s'intensifier »[2].

Le panel longitudinal se comporte de manière bien différente : en retenant des internautes ayant eu une activité en 2000 et en 2002, nous sommes assurés de n'observer que des individus qui ont intégré l'usage d'Internet dans leurs pratiques. Pas d'abandonnistes dans ce panel, donc, ce que traduit le nombre relativement stable d'actifs dans le panel mois par mois (voir Figure 5.2). Les variations sont ici imputables aux rythmes saisonniers, avec en particulier des baisses d'activité en février et surtout en juillet-août. Le nombre de sessions est particulièrement influencé par ces variations saisonnières[3], mais il connaît une croissance globale nette sur l'ensemble des 34 mois d'observation. On voit ici une confirmation de l'hypothèse formulée dans [Beaudouin *et al.* 2002] : pour les internautes fidèles, l'intensification de la pratique semble globalement se poursuivre sur trois ans.

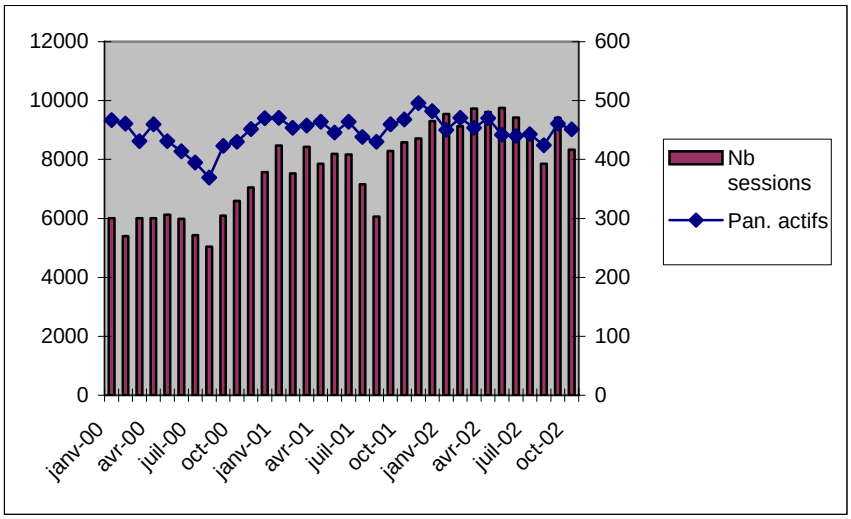


Figure 5.2. Activité du panel SN00-02

Comme nous avons pu le voir dans l'examen des usages globaux d'Internet, cette intensification globale n'est pas partagée par tous les individus. Pour quantifier ce phénomène, nous avons réalisé une régression linéaire du nombre et de la durée

---

[1] Voir Annexe 1 pour une description du projet TypWeb.

[2] [Beaudouin *et al.* 2002], p. 23.

[3] Il suffit, évidemment, d'une connexion dans le mois pour qu'un panéliste soit considéré comme actif, mais celui-ci peut être absent de son domicile pendant le reste du mois.

d'activité Web par mois pour chaque individu sur les 34 mois d'observation (voir Tableau 5.14).

Tableau 5.14. Pentes des régressions linéaires du nombre et de la durée des sessions par mois, SN00-02

		Nombre de sessions	Durée de sessions
Moyenne		0,206	15 min. 34 s.
Médiane		0,089	1 min. 22 s.
Quartiles	25	- 0,079	- 1 min. 37 s.
	50	0,089	1 min. 22 s.
	75	0,370	9 min. 50 s.

Pour la cohorte SN00-02, on observe en moyenne une session Web de plus par mois tous les cinq mois, et une durée d'activité Web augmentant de près de 15 minutes par mois. Ces valeurs moyennes sont fortement influencées à la hausse par une faible part des individus dont les pratiques se sont très fortement intensifiées : l'examen des valeurs médianes et des quartiles montre que, dans l'ensemble, le nombre de sessions augmente très lentement, environ une session mensuelle supplémentaire par an pour 1 min. 20 de navigation en plus par mois. En définitive, l'intensification de la pratique va plutôt dans le sens d'un allongement des durées de session (entre 1'20 et 9'50 minutes par mois pour un quart du panel, plus de 10 minutes pour un quart plus actif encore) que dans celui d'une augmentation du nombre de sessions, lequel ne connaît pas d'évolution significative, voire diminue faiblement pour un quart du panel. Des deux hypothèses que l'on opposait initialement, celle de l'intensification de la pratique n'est finalement vérifiée que pour une minorité des internautes ; pour les autres, l'ancrage de l'usage d'Internet dans les pratiques a conduit à une stabilisation globale du temps dévolu à cette activité.

Rythmes d'activité

A une échelle d'analyse différente, l'examen des rythmes d'activité dans la semaine et dans la journée permet de voir comment l'activité de navigation s'insère dans le cadre global des activités domestiques des trois panels résidentiels.

L'intensité d'usage varie au cours de la journée et de la semaine pour l'ensemble des trois sources de données, et ce de manière différente pour chaque panel : elle correspond à des rythmes d'activité globaux des utilisateurs. Pour les deux panels généralistes, les pics d'activité ont lieu le lundi et le mercredi (voir Figure 5.3 pour le panel SN2002). La cohorte suivie sur 34 mois se distingue uniquement par une activité particulièrement intense également le dimanche, en nombre de sessions comme en nombre de panélistes, tandis que l'échantillon SN2002 connaît peu d'activité ce jour.

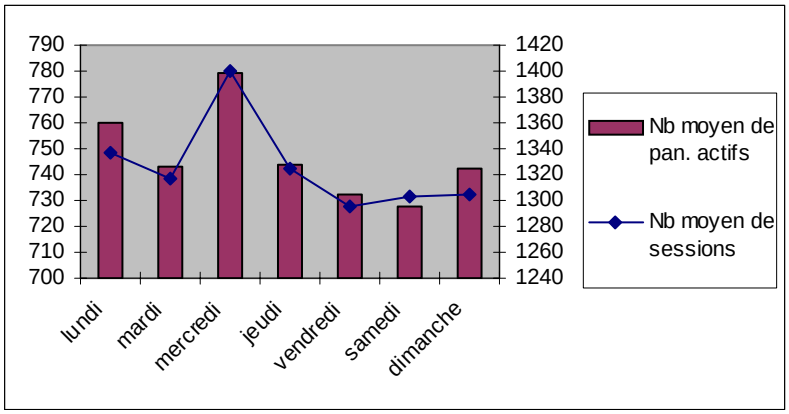


Figure 5.3. SN2002 – Activité par jour de la semaine

La répartition des heures d'activité montre une opposition bien plus nette entre semaine et week-end, relative au rythme de journées de travail. Pour les deux panels SensNet, les comportements sont similaires (voir Figure 5.4 pour les données SN00-02) : du lundi au vendredi, un pic d'activité très net se produit entre 18 et 23 heures, c'est-à-dire globalement à l'heure du retour au domicile. Le samedi et le dimanche, l'activité est bien plus répartie entre 10 heures et 22 heures, avec un creux d'activité vers 13-14 heures, et une hausse aux alentours de 17-18 heures, les soirées étant bien moins investies que le reste de la semaine.

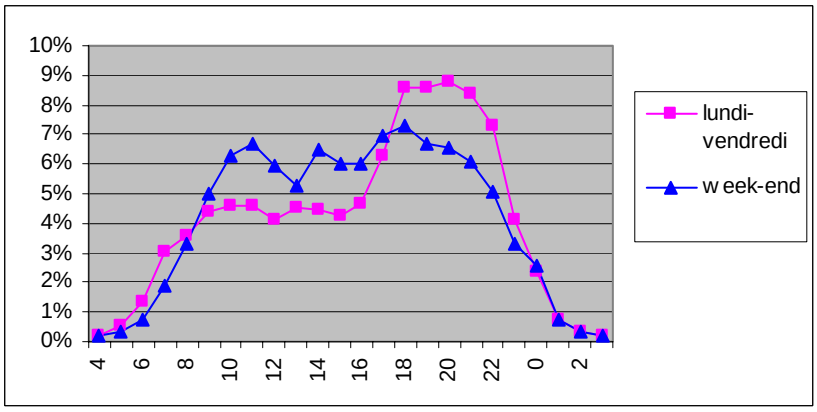


Figure 5.4. SN00-02 – Part des sessions par heure de début, semaine/week-end

Le profil du panel BibUsages est assez différent des deux autres : ce panel est mixte professionnel/résidentiel[1] et comporte un quart de non-actifs (retraités et étudiants en particulier), et son emploi du temps est structuré différemment de celui des données SensNet. L'activité Web reflète cette spécificité : l'activité hebdomadaire

[1] La sonde de recueil de trafic est dans certains cas installée sur le lieu de travail, et à l'inverse, des panélistes exercent leur activité professionnelle à domicile.

est répartie bien plus également entre les différentes journées. Le nombre de sessions est plus important du lundi au vendredi que le week-end, avec 30 à 32 panélistes actifs par jour contre 28 en moyenne. En ce qui concerne la répartition horaire, on ne constate aucune différence significative entre semaine et week-end : les sessions sont réalisées surtout entre huit heures le matin et 10 heures le soir, et leur nombre est assez stable entre ces deux horaires, malgré quelques pics d'activité à 11 heures, 14 heures et 20 heures.

Durée des sessions

L'activité de navigation s'insère donc fort naturellement dans la gestion globale du temps au quotidien et la disponibilité des utilisateurs. En revanche, elle n'est pas déterminée intrinsèquement par cette rythmique journalière : la durée des sessions est similaire pour les trois jeux de données, entre 30 et 35 minutes en moyenne (voir Tableau 5.15).

Tableau 5.15. Durées de sessions

	BibUsages	SN00-02	SN2002
Moyenne	30 min. 22 sec.	33 min. 57 sec.	35 min. 10 sec.
Médiane	13 min. 12 sec.	14 min. 21 sec.	14 min. 13 sec.
1er / 2e quartile	3 min. 7 sec.	4 min. 15 sec.	3 min. 48 sec.
2e / 3e quartile	13 min. 12 sec.	14 min. 21 sec.	14 min. 13 sec.
3e / 4e quartile	36 min. 57 sec.	35 min. 14 sec.	36 min. 43 sec.

Ainsi, malgré des activités de navigation de nature très différentes entre les trois panels (en particulier pour BibUsages), une séquence d'activité sur le Web est, dans sa durée, indépendante des éléments externes (individus, centres d'intérêt globaux) qui les composent, et la composante temporelle est bien plutôt liée à des éléments intrinsèques à la session. Pour les trois jeux de données, on trouve un premier quart des sessions très courtes, de moins de 3 à 4 minutes, tandis que la durée médiane se place autour d'un peu moins d'un quart d'heure ; lorsque la session s'allonge, la demi-heure est une limite dépassée dans un quart des cas uniquement, mais la durée peut alors être très importante, et atteint une heure en moyenne.

La gestion de leur « crédit-temps » par les utilisateurs est un de ces éléments : la grande diversité des sessions en termes de durée est liée à l'heure de début de session. Les quatre groupes de sessions – très courtes, courtes, longues, très longues – ne sont pas répartis de la même manière au fil de la journée (voir Figure 5.5 ci-dessous) ; nous travaillons ici sur les données SN2002, les plus représentatives et les plus récentes de nos trois jeux de données. On remarque que les sessions de plus de 35 minutes ne sont majoritaires qu'entre 14 heures et 17 heures, puis entre 20 et 22 heures, avec un pic particulier pour les sessions commençant entre 21 et 22 heures, et entre minuit et 1 heure du matin. Les autres tranches horaires laissent place aux sessions très courtes, mais surtout aux sessions à la durée comprise entre 4 minutes et un quart d'heure (2^{e} quartile).

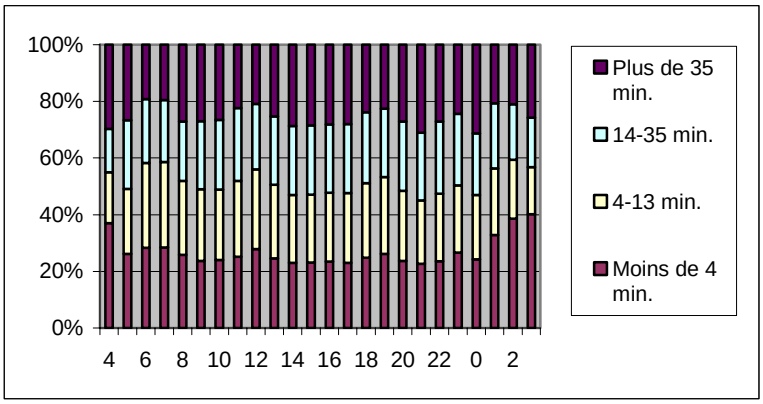


Figure 5.5. Durée des sessions et heure de début de session – SN2002, lundi-vendredi

Autre élément notable, on a constaté que les distributions sont très similaires entre le week-end et le reste de la semaine : alors que l'activité en nombre de sessions connaît des variations importantes entre ces deux moments de la semaine, la durée des sessions reste globalement très déterminée par l'heure de début de session quel que soit le jour de la semaine.

Il y a donc des périodes dans la journée où les sessions sont plus courtes, correspondant aux heures de retour du travail et d'avant-repas, dont on peut supposer qu'elles sont dévolues à certaines activités particulières sur le Web. Nous savons par ailleurs, par le biais d'études sur les modes de vie, que ces moments de la journée sont soumis à une nécessité particulière d'arbitrages entre un nombre important de sollicitations dans une période courte. À l'inverse, à partir de 22 heures, la disponibilité semble plus importante, et les sessions s'allongent. L'examen du contenu des sessions permettra par la suite de voir si ces sessions courtes sont liées à des tâches particulières ; contentons-nous pour l'heure de noter que la navigation sur le Web est une activité qui entre en concurrence avec d'autres au sein de la journée, et que sa pratique entre pour chaque individu dans l'économie générale du « crédit-temps ».

> ***Synthèse.*** *L'analyse, au sein des données volumineuses de trafic des trois panels, de la répartition de l'activité de navigation dans la semaine et dans la journée montre son inscription globale dans les pratiques dans et hors Web. Ce contexte global est fortement lié à la durée des sessions, et dénote l'implication d'activités différenciées au sein des sessions.*

5.2.2 Rythmes et formes de parcours

Nous avons jusqu'alors envisagé les sessions comme des blocs d'activité ; pour aller plus avant, nous entrons maintenant à l'intérieur des sessions, par le biais des indicateurs topologiques que nous avons élaborés (voir chapitre 4.2, « Analyser la séquentialité »). Nous introduisons ici le nombre de sites et de pages visités, le temps passé sur chaque page et chaque site, et l'enchaînement séquentiel des sites visités dans les sessions.

Sites visités

Le nombre de sites visités dans les sessions connaît des variations similaires à celles observées sur la durée, et les recoupe globalement. En moyenne, selon le jeu de données que nous manipulons, une session conduit à visiter entre 6 et 9 sites (voir Tableau 5.16), les sessions BibUsages se distinguant par un nombre plus important de sites que pour les deux autres échantillons. Mais comme pour la durée des sessions, la moyenne masque une importante diversité : la division en quartiles montre qu'un quart des sessions n'amène à visiter qu'un ou deux sites, tandis que seule la moitié des sessions entraîne la visite de plus de cinq sites. Pour le panel BibUsages, la moyenne de sites visités dans une session est plus importante du fait d'un plus grand nombre de sessions avec beaucoup de sites, les limites entre 1er / 2e et 2e / 3e quartiles étant similaires pour les trois jeux de données.

Tableau 5.16. Nombre de sites/portails visités par session

	BibUsages	SN00-02	SN2002
Moyenne	8,35	6,94	6,53
Médiane	4	4	4
1er / 2e quartile	2	2	2
2e / 3e quartile	4	4	4
3e / 4e quartile	9	8	7

Cette variation du nombre de sites visités dans chaque session est liée à celle déjà observée pour leur durée dans le cadre du rythme hebdomadaire. Le nombre de sites visités dans chaque session par période de la semaine ne connaît pas de différence notable entre le week-end et les autres jours. En revanche, la confrontation du nombre de sites visités et des rythmes d'activité journaliers déjà observés pour la durée des sessions est plus productive : pour les données SN2002, du lundi au vendredi, le nombre de sessions de plus de cinq sites diminue dans les tranches 12 h – 13 h et 19 h – 20 h, tandis que les autres sessions restent constantes à la même heure. Inversement, entre 20 h et 21 h, les sessions avec peu de sites décroissent, tandis que les autres connaissent un pic. Le week-end, cette différence de rythme entre sessions contenant un à quatre sites et sessions de plus de cinq sites est également observable : pic vers midi et 19 heures pour les premières, vers 15 heures et 18 heures pour les secondes.

Pour les données BibUsages, dont nous avons vu qu'elles ne comportent pas de différence marquée de rythme d'activité entre week-end et reste de la semaine, on retrouve des éléments similaires, mais avec des seuils différents. Alors que le panel généraliste opposait les sessions autour du seuil de cinq sites visités, les comportements sont ici similaires jusqu'à dix sites. C'est au-delà de dix sites visités que les décalages journaliers sont observables, ce qui semble dénoter des différences de pratique. Nous avons déjà observé que pour les panélistes de BibUsages, la durée médiane des sessions est, heure par heure, globalement inférieure à celle observée pour les deux autres panels ; il apparaît ici que, replongées dans le cadre des pratiques quotidiennes et de leur rythme, une session BibUsages « rapide » effectuée à des heures où la navigation entre en concurrence forte avec d'autres activités, dure à la fois moins longtemps et amène à voir plus de sites, comme si la navigation était accélérée, plus ciblée.

Durée de session et nombre de sites visités dans la session sont étroitement liés, c'est une évidence ; pour autant, la précédente comparaison entre données BibUsages et SN2002 invite à examiner plus finement cette corrélation. Si dans un temps donné, il semble bien difficile de voir plus d'un certain nombre de sites, en particulier pour les sessions courtes, il est intéressant de voir si ce lien est conservé pour l'ensemble des sessions, en croisant durées des sessions et nombre de sites visités (voir Tableau 5.17).

Tableau 5.17. Répartition des durées de sessions par nombre de sites visités

		1-2 sites	3-4 sites	5-10 sites	Plus de 10 sites
BibUsages	0-3 min.	67,2 %	27,8 %	7,4 %	0,6 %
	4-14 min.	18,5 %	37,6 %	30,3 %	7,9 %
	15-34 min.	9,1 %	22,6 %	33,4 %	26,3 %
	35 min. et plus	5,3 %	12,0 %	28,9 %	65,2 %
SN2002	0-3 min.	57,2 %	19,1 %	3,1 %	0,2 %
	4-14 min.	24,9 %	39,7 %	24,5 %	4,2 %
	15-34 min.	12,2 %	26,8 %	37,4 %	22,5 %
	35 min. et plus	5,6 %	14,4 %	35,0 %	73,1 %
SN00-02	0-3 min.	58,8 %	18,6 %	2,9 %	0,3 %
	4-14 min.	25,5 %	40,2 %	25,1 %	4,8 %
	15-34 min.	11,2 %	27,6 %	40,0 %	26,2 %
	35 min. et plus	4,5 %	13,6 %	32,0 %	68,8 %

Clef de lecture : dans les données BibUsages, 67,2 % des sessions comprenant 1 ou 2 sites durent moins de 4 minutes, 18,5 % durent entre 4 et 14 minutes.

Les différences entre jeux de données semblent attester un important effet d'ancienneté de la pratique : plus les internautes sont anciens et « avertis », plus ils voient de sites dans un temps donné. Les sessions d'un ou deux sites durent moins de 4 minutes dans 67 % des cas pour les données BibUsages, contre 57 % pour le panel généraliste SN2002. L'effet de rapidité est confirmé dans les sessions de plus de 10 sites : 65,2 % de ces sessions durent plus de 35 minutes pour BibUsages, contre 73,1 % pour SN2002. Dans tous les cas, le panel SN00-02 a une position intermédiaire.

Si l'on travaille à l'échelle de la page, on observe plus finement ce phénomène, ce que présente la Figure 5.6 (nous ne présentons pas le calcul pour les données SN00-02, similaire à celui pour SN2002). À mesure que le nombre de pages vues dans les sessions augmente, la durée des sessions croît régulièrement, avec un pic pour le dernier décile (valeurs extrêmes ou aberrantes). Parallèlement, pour BibUsages, le temps passé sur chaque page vue est relativement constant pour la première moitié des sessions, puis décroît de 34 à 20 secondes par page sur les 5 derniers déciles. L'effet « d'accélération » n'est ainsi perceptible que pour les sessions les plus longues, le rythme de visualisation des pages étant relativement constant pour les autres sessions.

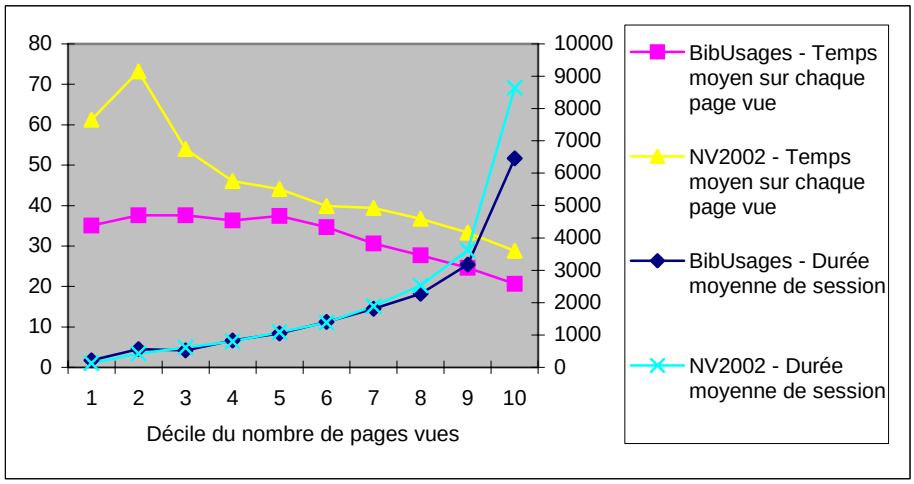


Figure 5.6. Nombre de pages vues et durées (BibUsages et SN2002)

Les sessions de SN2002 connaissent elles un effet d'accélération constant au fur et à mesure que l'utilisateur voit plus de pages dans la session. Il semble que les panélistes SensNet passent plus de temps pour appréhender le contenu des pages, possible reflet de l'effet d'apprentissage et d'expertise : il est probable que les utilisateurs de BibUsages, plus anciens et plus expérimentés, s'orientent plus facilement et plus rapidement à l'intérieur des sites et des pages Web.

Linéarité des parcours

L'examen des indicateurs topologiques permet d'aller plus loin dans l'analyse, en introduisant les notions de détour, de revisite de sites et de pages et de linéarité. Dans quelle mesure les sessions sont-elles linéaires, et si elles ne le sont pas, quelle est la portée des retours au sein des sessions, et quelles différences dans les modes de navigation cela dénote-t-il ? Comme nous l'avons présenté au Chapitre 4, nous distinguons deux types de linéarité : celle, globale, se situant au niveau de chaque requête adressée par le panéliste, dite « inter-pages », et une autre où les URL sont regroupées par site/portail, dite « inter-sites ».

Tableau 5.18. Part des sessions linéaires dans l'ensemble des sessions

	BibUsages	SN00-02	SN2002
Sessions inter-pages	24,5 %	17,1 %	18,6 %
Sessions inter-sites	31,3 %	35,3 %	36 %

Les deux modes de calcul apportent des résultats différents selon les jeux de données (voir Tableau 5.18). Pour le panel BibUsages, les sessions linéaires représentent environ un quart de l'ensemble des sessions dans les deux cas ; pour les deux autres panels, l'écart est important entre les linéarités inter-pages, qui concerne près de 18 % des sessions, et inter-sites, représentant près d'un tiers des sessions. Ainsi, si les retours sur des sites déjà vus au sein d'un parcours touchent globalement un tiers des sessions, à l'intérieur des sites, il semble que les utilisateurs de BibUsages aient des trajectoires sensiblement plus rectilignes que les autres.

Ici encore, on peut y lire un effet d'ancienneté de la pratique : le panel SN2002, où la proportion de nouveaux et récents internautes est la plus forte, est celui dont les sessions inter-pages sont les moins directes. En revanche, l'intensité de la pratique n'est pas liée à ce phénomène : nous avons examiné dans les données SN2002 la part des sessions linéaires en fonction du nombre total de sessions réalisées par panéliste sur les 10 mois d'observation. Au terme de ce calcul, aucune différence notable entre les d'utilisateurs : dans l'ensemble, pour chaque panéliste, la part des sessions linéaires inter-pages oscille entre autour de 29 %, celle des sessions linéaires inter-sites entre 31 et 36 %.

On ne s'étonnera pas de ce que les sessions linéaires sont plus courtes et comportent moins de sites que les autres : plus on voit de sites dans une session, plus celle-ci s'allonge, et plus la probabilité de revenir sur un site est importante. Ceci se vérifie pour les trois jeux de données que ce soit avec les sessions inter-pages (voir Tableau 5.19) ou les sessions inter-sites (voir Tableau 5.20).

Tableau 5.19. Sessions inter-pages linéaires et non linéaires (moyennes et médianes)

		Nb URL distinctes	Nb sites distincts	Durée totale	Temps par page vue
Sessions linéaires	BibUsages	7,1 (5)	2,6 (2)	2'43 (1'05)	32 (19)
	SN00-02	5,5 (3)	2,2 (2)	4'34 (1'05)	51 (29)
	SN2002	5,4 (3)	2,1 (1)	4'45 (1'09)	52 (30)
Sessions non linéaires	BibUsages	68 (36)	11,8 (7)	39'01 (21'23)	32 (21)
	SN00-02	40,4 (25)	8,8 (6)	39'40 (18'52)	42 (24)
	SN2002	42,6 (25)	8,5 (5)	41'19 (19'07)	43 (24)

De page en page, les temps moyen et médian passés sur chaque page sont similaires dans les sessions linéaires et non linéaires pour les sessions BibUsages, tandis que la revisite de pages amènent les panélistes NetValue à passer moins de temps sur chaque page et, en quelque sorte, à accélérer le rythme de la navigation.

Tableau 5.20. Sessions inter-sites linéaires et non linéaires (moyennes et médianes)

		Nb de sites distincts	Nombre de sites vus	Durée totale	Temps par site vu
Sessions linéaires	BibUsages	1,8 (1)	1,8 (1)	7'37 (1'44)	4'34 (1'07)
	SN00-02	2 (2)	2 (2)	8'48 (3'05)	4'57 (1'34)
	SN2002	1,9 (1)	1,9 (1)	9'22 (2'45)	5'40 (1'39)
Sessions non linéaires	BibUsages	11,3 (7)	35 (16)	40'44 (22'42)	2'01 (1'07)
	SN00-02	9,6 (6)	25,4 (13)	47'35 (24'34)	2'49 (1'30)
	SN2002	9,1 (6)	28,4 (13)	49'42 (25'00)	2'48 (1'33)

Ce phénomène est également observable à l'échelle des sites (voir Tableau 5.20) : dans les sessions non linéaires, le nombre de sites différents visités est sensiblement plus important (médianes à 7 ou 8, contre 1 ou 2 pour les parcours directs). On notera que les valeurs médianes du temps passé sur chaque visite de site sont systématiquement inférieures aux valeurs moyennes, mais surtout que l'écart entre moyenne et médiane est bien plus important pour les sessions linéaires : dans ce cas, quelques séquences de très longue durée influencent le calcul de la moyenne, valeurs

extrêmes bien moins présentes dans les sessions non linéaires[1]. Enfin, ici encore, le trafic BibUsages se démarque des deux autres par des durées plus faibles des séquences sur les sites et des sessions en général.

Ces éléments confirment la triple corrélation : nombre de sites différents / durée / linéarité. Nous avons déjà observé une relation quasi-linéaire entre durée de sessions et nombre de sites distincts vus dans la session ; l'examen de la proportion de sessions linéaires par décile de la durée de la session et par décile du nombre de sites visités montre une relation tout autre. À l'échelle de la page, plus la session dure, plus la part des sessions linéaires diminue, de manière proportionnelle ; à l'échelle du site, la distribution n'est pas la même, et prend une allure plus zipfienne.

Les résultats sont très similaires pour les trois échantillons, et restent stables si l'on croise linéarité et nombre de sites visités. Dans tous les cas, les sessions courtes, comportant peu de sites (s'il n'y en a qu'un, tout est joué d'avance) sont linéaires à 80 % en inter-pages et 92 % en inter-sites ; ces proportions décroissent très vite à mesure que la session s'allonge, et au septième décile, 3 % sont linéaires en inter-pages, et 12 % en inter-sites. Ainsi, au-delà de quatre sites différents visités dans une session et de dix minutes de navigation, les deux tiers des sessions comportent au moins un retour sur un site déjà visité dans la session, et 90 % amènent à voir une même page au moins deux fois.

> ***Synthèse.*** *La diversité des sessions en termes de durée se retrouve dans le nombre de pages et de sites différents visités dans la session : à mesure que le nombre de pages vues dans les sessions augmente, leur durée croît régulièrement. Ces éléments sont corrélés à la linéarité des parcours : dès que la session s'allonge, la ligne brisée se généralise, à l'échelle de la page comme du site. Une première opposition se fait jour entre des sessions courtes et linéaires, qui regroupent un tiers des sessions à l'échelle du site, et semblent plus ciblées, et des sessions longues et non-linéaires reflétant des comportement plus diversifiés.*

Quantifier et qualifier les revisites à l'échelle de la page

Les sessions non linéaires sont donc relativement fréquentes dans nos corpus de sessions, et font partie intégrante des modes de navigation sur le Web puisqu'elles concernent entre 75 et 80 % des sessions. Toutefois, l'opposition linéaire/non linéaire masque une grande diversité au sein des sessions non linéaires, qu'il importe d'examiner plus en détail. Entre le simple retour sur une page dans une longue session quasi-rectiligne et la visite massive et répétée d'une même page dans une navigation en étoile, par exemple la page de réponse d'un moteur de recherche, ce sont différents modes de navigation qui sont en jeux. L'examen des taux de linéarité calculés pour chaque session permet d'apprécier cette diversité, en donnant une

---

[1] Il n'est pas impossible que ces très longues séquences linéaires ne soient pas la trace d'une navigation volontaire et supervisée de la part des internautes, mais plutôt de requêtes adressées automatiquement (rafraîchissement automatique de pages, bandeaux publicitaires alternés, etc.).

estimation de la part de la session consacrée à des pages ou des sites déjà vus, en nombre de pages et de sites ainsi qu'en durée.

Tableau 5.21. Valeurs des taux de linéarité r_{page} *et* r_{site} *par corpus (sessions non linéaires)*

	Sessions inter-pages non linéaires		Sessions inter-sites non linéaires	
	Moyenne	Médiane	Moyenne	Médiane
BibUsages	0,69	0,75	0,47	0,47
SN00-02	0,63	0,66	0,50	0,50
SN2002	0,63	0,67	0,48	0,50

Les valeurs moyennes des taux de linéarité r_{page} et r_{site} pour les sessions non linéaires sont relativement comparables pour les trois jeux de données : en moyenne, la linéarité à l'échelle de la page est de l'ordre de 0,65, et de 0,5 à l'échelle du site (voir Tableau 5.21). La distribution des valeurs de r_{page} montre une concentration autour de la moyenne (proche de la médiane)[1]. En d'autres termes, dès lors que l'utilisateur revisite des pages, un tiers des pas de la session correspond à des pages vues plusieurs fois en moyenne, et ce pour les trois jeux de données. Tout au plus notera-t-on que pour les sessions du panel BibUsages, la linéarité est globalement plus importante au niveau de la page, mais l'est moins au niveau du site.

Le taux de linéarité calculé sur la base de la durée de visite des pages, d_{page}, rend compte du temps passé sur les pages vues une fois dans la session : il permet d'analyser plus finement la revisite en la quantifiant sur une échelle de temps. L'indicateur montre qu'en moyenne la moitié du temps de la session concerne des revisites, avec des taux avoisinant 0,5 (voir Tableau 5.22).

Tableau 5.22. Valeurs de d_{page} *pour les sessions inter-pages non linéaires*

		BibUsages	SN00-02	SN2002
Moyenne		0,59	0,53	0,53
Médiane		0,65	0,55	0,55
Quartiles	25	0,40	0,32	0,32
	50	0,65	0,55	0,55
	75	0,82	0,76	0,76

Clef de lecture : en moyenne, pour le panel BibUsages, 59 % de la durée d'une session non linéaire sont consacrés à des pages vues une seule fois dans la session.

La distribution des valeurs du taux d_{page} calculé sur la durée est beaucoup plus uniforme que celle observée avec le calcul basé sur les éléments visités (voir Figure 5.7 ci-dessous) : le pic autour de la moyenne est moins sensible, et une continuité forte s'établit entre sessions linéaires et sessions non linéaires.

---

[1] La distribution du taux de répétition moyen r_{page}, pour les sessions non linéaires montre un pic autour de la valeur 0,75, et un déficit de valeurs entre 0,75 et 1 (sessions linéaires). Nous pensons que cette distribution ne correspond pas tant à une rupture entre sessions linéaires et sessions non linéaires qu'à un biais issu du mode de calcul de l'indicateur : en effet, le nombre important de sessions dont le nombre de pages vues est de trois, quatre ou un multiple faible de ces valeurs (8, 9, 12), favorise des ratios de 0,66 ou 0,75 lorsqu'une page est revue sur trois par exemple.

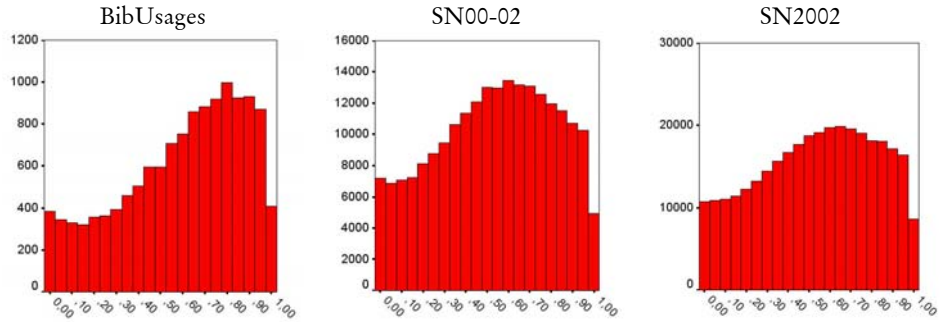


Figure 5.7. Fréquences des valeurs de d_{page} *pour les sessions inter-pages non linéaires*

On observe une différence sensible entre les sessions BibUsages et les autres : le taux de linéarité est globalement plus élevé, ce qui signifie que les utilisateurs de ce panel passent moins de temps sur les pages vues plus d'une fois. Ce constat, qui rejoint celui d'un rythme plus élevé dans la navigation, tend à montrer que ces internautes sont plus sélectifs et plus rythmés dans leurs visites. Cet effet est confirmé par le profil intermédiaire des données SN00-02 entre BibUsages et SN2002, panel intermédiaire en termes d'ancienneté et d'expérience.

Comment s'organisent structurellement les revisites à l'échelle de la page ? Touchent-elles certaines pages en particulier, ou s'appliquent-elles à l'ensemble des pages vues plusieurs fois dans la session ? Le taux de concentration c_{page} qui rend compte de ces phénomènes est égal à 1 dans un quart des cas (chaque page revisitée est revue une seule fois), et reste globalement compris entre 1 et 2. Toutefois, les taux de concentrations élevés pour le dernier quart montrent un écart de comportement entre des sessions avec pages-pivots (concentration forte) et d'autres où la revisite est plus étalée sur les différentes pages vues (voir Tableau 5.23).

Tableau 5.23. Taux de concentration c_{page}

		BibUsages	SN00-02	SN2002
Moyenne		3,35	2,34	2,35
Médiane		1,65	1,57	1,61
Quartiles	25	1	1	1
	50	1,65	1,57	1,61
	75	2,54	2,23	2,36

Clef de lecture : pour le panel BibUsages, une page revisitée est revue en moyenne 3,35 fois dans la session.

Cet écart entre deux profils structurels de revisite est fortement lié au taux de linéarité : moins les sessions sont linéaires, plus les revisites se concentrent sur une ou plusieurs pages. Nous pouvons supposer que cet écart entre sessions peu linéaires, avec un fort taux de concentration, et sessions très linéaires est le reflet de stratégies et de configurations de navigation différentes : soit l'on est dans une configuration de « prédateur » où l'utilisateur sait précisément où il va et agit rapidement, soit l'on est dans une forme de parcours de type « découverte » ou « multi-tâche », où

certaines pages servent de pivot à la revisite ou au passage à une tâche différente (pages de résultats de moteurs de recherche, page d'accueil d'un site, page de démarrage, etc.).

Pour appuyer cette hypothèse, on examinera la présence des actions de type *back* dans les sessions non linéaires : le recours à cette fonctionnalité est en effet particulièrement fréquent dans des configurations avec pages-pivot où l'utilisateur explore des liens à partir d'une page donnée.

Dans les deux jeux de données SensNet, les comportements sont similaires : au sein des sessions non linéaires, 85 % comportent au moins une séquence de type *back*, quelle qu'en soit la longueur. Pour les données BibUsages, seules 70 % des sessions non linéaires sont concernées. Dans les trois cas, la longueur des séquences de ce type est majoritairement de 2 pas (75 % des cas), ce qui signifie que les trois quarts des mouvements de retour arrière ne concerne qu'une page et une seule, sous forme d'aller-retour sans profondeur.

La part des pages vues au sein d'actions de type *back* dans l'ensemble des parcours reste elle-même souvent modeste : l'indicateur b_{page} qui en rend compte vaut en moyenne 0,18 (médiane : 0,11) pour les sessions BibUsages, et 0,24 pour les sessions SensNet (médiane : 0,2). Ceci signifie que les pages au sein d'actions *back* occupent en moyenne entre 20 et 25 % des pages vues dans une session non linéaire. Les distributions des valeurs de l'indicateur renforcent ce constat, et montrent une fois de plus la spécificité du panel BibUsages, pour lequel les mouvements de *back* sont non seulement moins fréquents, mais occupent moins de place dans les parcours.

Ainsi, de manière générale, la ligne brisée est loin d'être une constante dans la navigation lorsqu'on l'examine au niveau de la page. Si les sessions non linéaires s'opposent aux sessions linéaires en termes de durées et de nombre de pages vues, elles forment un groupe globalement homogène où les revisites sont plutôt de d'ordre du « pas de côté » et du petit détour. Seule une minorité de sessions donne lieu à des comportement plus particuliers, avec des pages-pivot, des revisites intensives et de nombreux *back*, relevant de contextes de navigation spécifiques que l'analyse des contenus permettra d'éclairer.

Revisites de site en site

Les comportements de revisite examinés à l'échelle du site viennent confirmer ces hypothèses et les complètent. Nous avions vu qu'à l'échelle du site, les sessions sont plus linéaires qu'à celle de la page, avec un tiers de sessions linéaires en inter-sites contre 18 à 25 % en inter-pages. Par contre, au sein des sessions non linéaires, le taux de linéarité r_{site} est plus faible qu'à l'échelle de la page, avec une moyenne autour de 0,5 pour les trois jeux de données, identique à la médiane. La distribution des valeurs de r_{site} (Figure 5.8 ci-dessous) montre en effet une concentration du taux de linéarité autour de 0,5 ainsi qu'une rupture forte avec les sessions linéaires ($r_{site} = 1$).

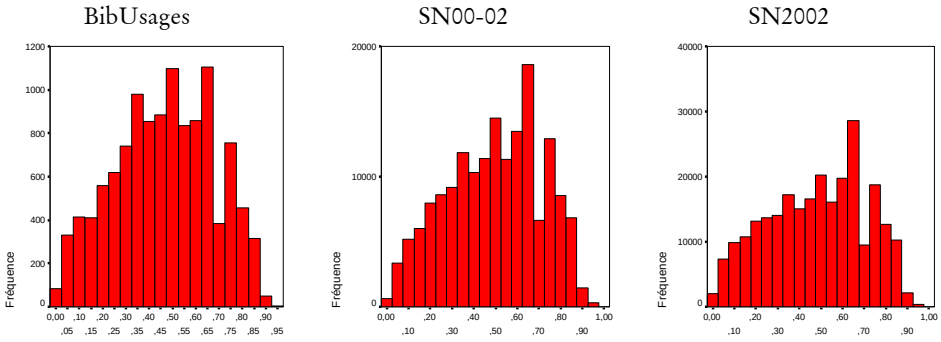


Figure 5.8. Fréquences des valeurs de r_{site} *pour les sessions inter-site non linéaires*

Le même indicateur calculé sur les durées de visite confirme l'importance des sites vus plusieurs fois dans les sessions : avec une valeur moyenne proche de 0,3 (médiane autour de 0,2), les sites vus plusieurs fois occupent en moyenne plus des deux tiers de la durée des sessions (voir Tableau 5.24).

Tableau 5.24. Valeurs de d_{site} *pour les sessions inter-sites non linéaires*

		BibUsages	SN00-02	SN2002
Moyenne		0,29	0,32	0,30
Médiane		0,19	0,24	0,21
Quartiles	25	0,04	0,06	0,04
	50	0,19	0,24	0,21
	75	0,47	0,53	0,52

La distribution des valeurs de d_{site} calculé sur la durée de visite confirme cette tendance, avec un contingent majoritaire de sessions où les sites vus une seule fois dans le parcours comptent pour moins de 10 % du temps total de la session. Le détail de l'activité sur ces sites vus plus d'une fois montre que les retours ne sont pas un simple passage rapide, mais correspondent au contraire à la majorité du temps passé sur le site revu. En effet, lorsqu'un site est vu plusieurs fois dans une session, le premier passage sur le site occupe en moyenne entre 26 à 30 % du temps total sur ce site selon le jeu de données considéré (médianes entre 18 et 24 %), avec une distribution orientée vers les faibles valeurs.

Revoir un site ne semble donc pas dû au hasard ni aux nécessités du cheminement dans les hyperliens, mais apparaît bien plus comme le signe d'un intérêt particulier pour les contenus proposés. Toutefois, revisiter un site n'amène globalement pas à le revoir plus de deux fois : le taux de concentration c_{site} reste en-deçà de 2 pour la moitié des sessions non linéaires à l'échelle du site, et ne dépasse 3,5 que pour un quart de ces sessions (voir Tableau 5.25).

Tableau 5.25. Taux de concentration c_{site}

		BibUsages	SN00-02	SN2002
Moyenne		3,22	2,81	3,31
Médiane		2	2	2
Minimum		1	1	1
Maximum		100	299	416
Quartiles	25	1,20	1	1
	50	2	2	2
	75	3,5	3,2	3,5

Les mouvements de retour arrière sont très fréquents dans la revisite de site : dans 95 à 96 % des sessions non linéaires, on observe des comportements de ce type. La part de ces retours arrière dans les parcours est bien plus importante qu'à l'échelle de la page : ils occupent en moyenne la moitié des pas de site en site dans les sessions non linéaires (voir Tableau 5.26), et la distribution des valeurs de b_{site} montre la concentration des valeurs autour cette moyenne.

Tableau 5.26. Valeurs de b_{site} *(BibUsages, SN00-02, SN2002)*

		BibUsages	SN00-02	SN2002
Moyenne		0,54	0,54	0,56
Médiane		0,53	0,54	0,56
Quartiles	25	0,40	0,40	0,40
	50	0,53	0,54	0,56
	75	0,67	0,67	0,71

Si l'on ôte à l'échelle du site les mouvements de *back* des sessions non linéaires, les sessions apparaissent comme linéaire dans 39,8 % des cas pour BibUsages, 43,9 % des cas pour SN00-02 et 46,6 % des cas pour SN2002. Cela signifie que dans plus de la moitié des cas, la non-linéarité est due à un mouvement de retour vers un ou plusieurs sites vus précédemment. On est ainsi en droit de supposer que certains sites servent de pivot à la navigation, et que les parcours s'articulent autour de ce site.

> ***Synthèse.*** *Le groupe des sessions non linéaires masque une grande diversité dans la structure des revisites. En premier lieu, à l'échelle de la page, les revisites sont plutôt de d'ordre du « pas de côté » et du petit détour, mais plus la session s'allonge et se complexifie, plus on voit apparaître des phénomènes de page-pivot sur lesquelles se concentrent les revisites. Ces éléments sont encore plus marqués à l'échelle du site : revoir un site ne semble pas dû au hasard, mais apparaît comme le signe d'un intérêt particulier pour les contenus proposés. Ces phénomènes de pages ou sites-pivot semblent être le reflet de deux modes de navigation différents : le comportement « prédateur » limite les revisites et demeure plutôt rectiligne, avec quelques digressions, tandis que les parcours de type « découverte » ou « multi-tâche » s'appuient sur certaines pages ou certains sites pour déployer une navigation en étoile plus longue et plus complexe.*

Conclusion

L'examen des éléments temporels, rythmiques et topologiques des sessions montre qu'il n'existe pas une session-type, mais des types de sessions bien différenciés, aux profils topologiques spécifiques. La plupart des éléments topologiques et rythmiques sont liés : durée, nombre de pages et de sites et revisite entretiennent des relations étroites. Si les corrélations sont fortes entre les différents indicateurs que nous avons construits, elles ont des causes structurelles : l'hétérogénéité des sessions sur le plan du nombre d'unités qui les composent rend difficile leur comparaison, et biaise le calcul des indicateurs. Il n'est alors pas surprenant que la linéarité soit directement liée au nombre de pages ou de sites visités. L'analyse temporelle permet de nuancer ces constats et d'avoir une vue plus qualitative sur les revisites : elle montre notamment le temps important qu'occupent les pages vues plusieurs fois dans les sessions, et le fait qu'un deuxième passage n'est pas un survol, mais donne lieu à une véritable lecture. L'analyse en termes de durée met également à jour des éléments de rythmique, non observables par d'autres moyens : sur ce plan, le panel BibUsages s'oppose fortement aux autres, et s'impose comme un panel « expert » où les utilisateurs ciblent leur navigation et passent peu de temps en détours.

En définitive, les sessions linéaires forment un groupe distinct des autres, pour lequel les indicateurs topologiques ont des valeurs fixes, et semblent renvoyer à des contextes précis où l'internaute visite de manière ciblée certains sites dans une courte durée. Pour les autres sessions, le panorama est plutôt diversifié : certaines sessions comportent peu de détours et les à-côté sont globalement minoritaires, tandis que d'autres se démarquent par de faibles taux de linéarité et l'emploi plus massif des *back*, et semblent dénoter des structures de navigation particulières. Ces différents comportements de navigation correspondent à des contextes d'usage différents, et une première distinction se fait jour entre des parcours ciblés, plutôt linéaires, et des parcours plus ouverts et plus longs, dont la structure complexe est caractérisée par la présence de pages-pivot et de nombreux retour-arrière. Cette opposition recouvre, dans une certaine mesure, celles observées dans la répartition horaire de l'activité : l'activité Web s'inscrit dans le contexte plus général de l'activité journalière, et certains profils de sessions sont plus présents à certaines heures qu'à d'autres. Toutefois, ces éléments formels et temporels ne sauraient expliquer seuls ces différences de comportements dans les parcours ; l'analyse des contenus visités d'une part, et de l'inscription des visites dans le corpus et la navigation de chaque panéliste de l'autre, permettront de donner sens à ces formes de parcours en les insérant au sein des pratiques en contexte.

5.3 Contenus des parcours

Parallèlement à l'étude de la topologie des sessions, on souhaite avoir un panorama des contenus visités par les internautes, et éprouver ce faisant la solidité de nos descripteurs. Dans cette optique, nous évaluons dans un premier temps précisément la capacité des annuaires à décrire les parcours seuls, puis montrons que le recours complémentaire à *CatService* permet d'améliorer à la fois la couverture et

la qualité des descriptions. Nous proposons ainsi une première segmentation des parcours sur la base des contenus visités par les internautes.

5.3.1 Étendue des descriptions de contenu

La mobilisation des ressources exogènes que représentent les annuaires Web pour décrire le contenu des parcours des internautes est à la fois riche et complexe. Elle doit être validée en termes quantitatifs et qualitatifs, au regard des différents modes de structuration et de description des objets du Web qu'ils utilisent.

Choix des annuaires exploités

Nous avons déjà vu que le taux de couverture moyen du trafic global des panels représentatifs en 2000, 2001 et 2002 est de l'ordre d'un tiers pour chaque annuaire, ce qui nous autorise à tenter d'exploiter cette source de données pour décrire le contenu des parcours. Plus en détail, le Tableau 5.27 présente, pour chaque jeu de données, le taux de couverture moyen des sessions par les annuaires aspirés en 2002.

Tableau 5.27. Couverture moyenne des sessions par les annuaires 2002

	BibUsages		SN2002		SN00-02	
	Nb. URL	Durée	Nb. URL	Durée	Nb. URL	Durée
Looksmart	35,3 %	34,8 %	31,8 %	29,6 %	36,6 %	34,8 %
Lycos	22,3 %	22,3 %	21,0 %	21,1 %	26,2 %	25,8 %
MSN	33,7 %	33,2 %	29,7 %	28,0 %	35,8 %	34,4 %
Nomade	34,0 %	33,7 %	30,2 %	28,2 %	33,7 %	31,9 %
Open Dir.	24,4 %	23,3 %	20,7 %	18,5 %	25,8 %	23,8 %
Voila	35,9 %	35,9 %	31,7 %	29,9 %	34,1 %	32,5 %
Voila PP	1,4 %	1,3 %	1,1 %	1,2 %	1,7 %	1,9 %
Yahoo	30,4 %	29,6 %	27,5 %	25,8 %	33,1 %	31,4 %

Nous avions déjà observé que les annuaires s'adaptent à l'évolution du Web, mais qu'ils ne précèdent pas le trafic : les meilleurs taux de couverture des annuaires 2002 avec le trafic global des panels NetValue étaient obtenus en 2001. On ne s'étonnera pas, dans ces conditions, que les données SN00-02 soient mieux décrites par les annuaires que celles enregistrées en 2002 uniquement. Plus surprenants sont les taux de couverture pour le panel BibUsages, systématiquement supérieurs aux deux autres, alors que ces données sont de toutes les plus tardives. On peut supposer que ce résultat est le reflet du caractère très académique du panel, qui fréquente beaucoup de sites institutionnels ou renommés, lesquels sont bien mieux représentés dans les annuaires que les autres (nouveaux sites, sites perso, etc.).

Deuxième élément notable, les taux de couverture sont relativement semblables selon qu'on les calcule en nombre d'URL ou en durée ; tout au plus remarque-t-on un écart constant d'un à deux points en faveur de la couverture en nombre d'URL vues, ce qui tend à montrer que, globalement, les internautes passent moins de temps, au sein des sessions, sur les sites indexés par les annuaires que sur les autres.

Nous nous sommes attachés à montrer au Chapitre 3, lors de la description des données issues des annuaires, que les huit annuaires étudiés diffèrent profondément

en termes de pages et de sites indexés, dans la manière dont ils décrivent ces pages et ces sites, et surtout en ce qui concerne leurs structures. Nous avons conclu qu'il est impossible de mobiliser simultanément les huit sources, et qu'il est préférable de projeter chaque annuaire séparément sur les parcours. Cette approche a en outre l'avantage d'opérer comme un système de vérification des résultats, les conclusions obtenues avec chacun des annuaires devant être cohérentes entre elles, *modulo* leurs spécificités thématiques.

La description des parcours par les catégories d'annuaires nécessite cependant d'examiner de près chaque annuaire avant de l'exploiter. En effet, certains éléments structurels se montrent gênants, voire rédhibitoires : on souhaite, pour avoir des descriptions fiables et relativement tranchées, qu'une adresse ne soit pas indexée à de multiples endroits, et que les plans de classement respectent une cohérence thématique générale qui ne fasse pas apparaître de catégorie « fourre-tout ».

De ce point de vue, tous les annuaires ne sont pas éligibles pour décrire les parcours : Looksmart duplique des pans entiers de ses catégories pour faciliter la navigation, et une URL a de grandes chances de figurer sous plusieurs catégories de premier niveau. Yahoo marque géographiquement ou économiquement les sites indexés, et inscrit quasi-systématiquement chaque URL indexée sous les catégories « Exploration géographique » ou « Commerce et économie » en même temps que sous une des autres catégories de premier niveau. Voila recourt abondamment à la multi-indexation, avec un taux de répétition moyen des URL de 3,24, ce qui est susceptible de « bruiter » l'analyse. Les taux de couverture sont également à prendre en compte : Voila Pages Perso mis à part, Lycos et Open Directory sont les plus mal lotis, avec en moyenne des taux avoisinant les 23 %, alors que leurs concurrents sont dix points au-dessus.

Au terme de ce travail d'examen des sources, nous avons choisi de retenir trois annuaires : MSN France, Nomade et Yahoo France. Ces trois annuaires ont les meilleurs taux de couverture (de l'ordre de 32 % en moyenne, en durée de session), recourent peu à la multi-indexation des URL, et présentent des structures assez équilibrées entre leurs différentes catégories de premier niveau (voir Tableau 3.26, Tableau 3.27 et Tableau 3.31, p. 125). Enfin, ils répondent chacun à des logiques de classement différentes : localisation géographique pour Yahoo, angle économique pour MSN, catégorisation orientée « monde de l'Internet » pour Nomade. De ce point de vue, ces trois annuaires se complètent bien et permettent une validation croisée des résultats.

Toutes les sessions n'étant pas décrites complètement par les annuaires, nous avons pour chaque couple données-annuaire retenu les sessions décrites à plus de 50 % par l'annuaire considéré[1]. Comme le montre le Tableau 5.28, nous perdons globalement 70 à 75 % de notre corpus de sessions dans cette sélection.

---

[1] Ce taux de couverture est calculé sur la base de la durée passée sur des pages décrites par les annuaires par rapport à la durée de la session.

Tableau 5.28. Part de l'ensemble des sessions couvertes à plus de 50 % pour chaque couple annuaire-données

	BibUsages	SN00-02	SN2002
MSN	29,9 %	31,1 %	24,3 %
Nomade	31,2 %	28,4 %	24,4 %
Yahoo	26,5 %	28,2 %	22,4 %

Nous avons bien tenté de réduire cet écart : n'existe-t-il pas, pour chaque URL visitée non décrite par l'annuaire, une page ou un site similaire qui serait, lui, décrit par l'annuaire ? Pour tester cette hypothèse audacieuse, nous avons sélectionné un sous-échantillon test de 80 000 sites comportant des URL visitées non indexées et pour chacun d'eux, envoyé sur le moteur de recherche Google une requête du type « related », qui renvoie les sites et pages similaires pour une adresse donnée. Le résultat est bien pauvre : sur les 48 400 sites pour lesquels Google proposait un site similaire, seulement 10 260 étaient indexés dans les annuaires. Si l'on ajoute à ce faible gain les distorsions que cette méthode implique (comment connaître, dans chaque cas, le degré de similarité renvoyé ?), on comprendra que nous ayons finalement laissé de côté cette piste.

Descriptions combinées

En exploitant séparément Nomade, MSN et Yahoo, sommes-nous pour autant tirés d'affaire, et disposons-nous de données exploitables ? Nous avons tenté une première description des sessions Web couvertes à plus de 50 % par les catégories de premier niveau des trois annuaires retenus. Pour chaque page décrite par l'annuaire, nous avons noté la ou les catégories de premier niveau correspondantes ; nous regardons ensuite dans la session le temps passé sur chaque catégorie, et retenons la catégorie la plus représentée dans la session. Pour les trois annuaires, le résultat s'avère très bruité, et presque écrasé par la part très importante des portails généralistes, qui font souvent office de page de démarrage en tant que fournisseurs d'accès, et drainent une forte audience en général du fait de la diversité des contenus et services qu'ils proposent. Pour MSN, c'est la catégorie « Informatique – Internet » qui ressort loin devant toutes les autres (voir Tableau 5.29 ci-dessous) ; dans le cas de Nomade, « Espace B to B » prend le dessus (38,2 % des sessions) ; et pour Yahoo, c'est la catégorie « Commerce et économie » qui écrase les autres (64 %).

Tableau 5.29. SN2002 : répartition des sessions par catégorie MSN la plus forte dans la session

Catégorie	Part des sessions
Informatique – Internet	40,2 %
Finances - Bourse – Patrimoine	9,3 %
Jeux – Consoles	7,3 %
Entreprises	7,2 %
Infos – Météo	6,6 %
Arts - Culture – Média	6,4 %
Loisirs - Passions	6,4 %
Vie quotidienne - Société	4,9 %
Shopping	3,4 %
Sports	2,2 %
Savoir - Education	2 %
Voyages - Tourisme	1,4 %
Emploi, formation	1,3 %
Santé	0,9 %
Sciences - Techniques	0,5 %

Clef de lecture : dans 6,6 % des sessions, la catégorie de premier niveau de MSN « Infos – Météo » est celle sur laquelle l'internaute a passé le plus de temps dans la session.

Il est apparu indispensable de « casser » cette catégorie dominante qui regroupe des contenus et des services très diversifiés, et de pouvoir descendre plus finement dans l'analyse des parcours sur les portails généralistes. Pour cela, nous avons mobilisé les informations fournies par *CatService*, qui décrit précisément les différents services utilisés sur ces sites. Pour chaque page vue dans une session, nous avons construit une description basée sur les types de portails et, pour les portails généralistes, les types de services, suivant les règles suivantes :

1. si le site est identifié par *CatService*,
    a. si c'est un site personnel hébergé gratuitement par un fournisseur d'accès ou un portail (Wanadoo, Free, Yahoo, etc.), on ne retient pas la description *CatService* et on passe au cas n°2 ;
    b. si la page correspond à un service de moteur de recherche, de WebMail ou de *chat*, elle est identifiée comme telle, même si le service est fourni par un portail généraliste ;
    c. si la page se situe sur un portail généraliste, on retient la description au niveau du service utilisé : « portail généraliste – informations », « portail généraliste – page d'accueil », etc.
2. si le site n'est pas identifié par *CatService* ou si c'est un site personnel, on cherche dans l'annuaire si l'URL visitée correspond à une URL indexée, et l'on retient les catégories de premier niveau de l'annuaire qui décrivent cette URL.

L'utilisation de *CatService* apporte en outre une solution à un problème de couverture qui touche essentiellement les portails généralistes. Notre méthode de projection des annuaires sur les parcours est basée sur un appariement au niveau des URL, elle implique que l'URL de l'annuaire soit un sous-ensemble de celle visitée. Les annuaires indexant la plupart du temps des points d'entrée des sites, nous manquons les sites réparties sur plusieurs sous-domaines : par exemple, si l'adresse du site de Free, `http://www.free.fr` est indexée par un annuaire, nous manquerons dans les

parcours la visite des pages sur l'offre ADSL de Free, dont les adresses commencent par `http://adsl.free.fr`, ou encore le WebMail de Free, sous `http://imp.free.fr` et quelques autres variantes. Ce biais technique s'avère d'autant plus gênant que ce sont justement les sites de taille importante, au premier rang desquels les grands portails généralistes, qui recourent à ce dispositif technique. Le recours à *CatService*, qui couvre très bien ce problème de variation du *host*, permet de pallier ce problème.

Enfin, les annuaires laissent dans l'ombre des sites qui drainent de fortes audiences mais apparaissent peut-être comme moins « présentables » : sites pornographiques et sites tournant autour de la galaxie du piratage. De la même manière qu'à la grande époque du Minitel, les services de messagerie, de dialogue et de rencontre « roses » drainaient la majorité de l'audience, le Web « rose » fait partie intégrante des pratiques sans avoir de représentation correspondante dans les services de recherche généralistes grand public que constituent les annuaires Web ici utilisés.

Pour ne pas laisser de côté les sites dits « adultes », nous avons mobilisé une information fournie par NetValue dans les données SensNet, qui identifie ces sites ; nous ajoutons ainsi, aux côtés des catégories d'annuaires, une catégorie « pornographie ». Cette description additionnelle complète bien celles des annuaires : pour les données SN2002, seules 0,8 à 1,2 % des pages vues décrites par les annuaires le sont également par la catégorie « pornographie ».

Ces trois sources de descriptions se révèlent très complémentaires (voir Figure 5.9 pour MSN, et Figure 5.10 pour Nomade) : la catégorie « pornographie » s'avère très indépendante des deux autres descripteurs, tandis que les portails généralistes et les services de recherche et de communication par *CatService* (24,1 % des URL de SN2002) et annuaires (couvrant 21 à 24 % des URL SN2002) ne se recoupent que pour 4 à 5 % des URL des données SensNet 2002.

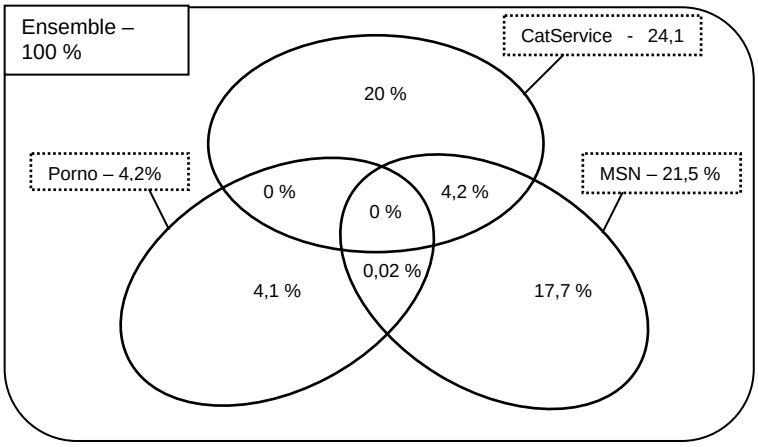


Figure 5.9. Couvertures croisées pour SN2002 (MSN, CatService, pornographique)

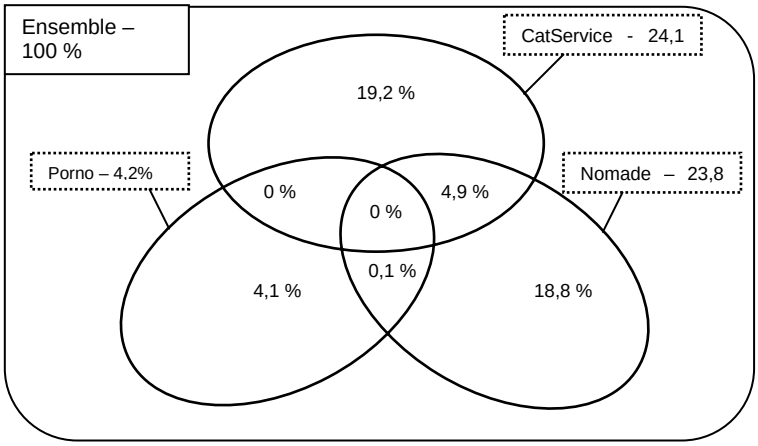


Figure 5.10. Couvertures croisées pour SN2002 (Nomade, CatService, pornographique)

Avec cette utilisation conjointe des annuaires, de *CatService* et de l'information fournie par NetValue sur les sites pornographiques, la liste des descripteurs de sessions contient une base, constituée des informations sur les quinze types de services pour les portails généralistes, des trois services de communication et de recherche et d'une étiquette « porno » (voir Tableau 5.30), à laquelle s'ajoutent les différentes catégories de premier niveau de chaque annuaire (voir Tableau 5.31).

Tableau 5.30. Catégories hors annuaires utilisées pour décrire les sessions

Services Web
Moteur WebMail WebChat
Portails généraliste : services
Achat Aide Annuaire Bourse Communication Divers Généralités Information Produit Information Service Informations Loisir En Ligne Non catégorisé Page Accueil Page Perso Personnalisation / Données Personnelles
Site pornographique

Tableau 5.31. Catégories de premier niveau pour les trois annuaires retenus

MSN (15 catégories)	Nomade (12 catégories)	Yahoo (14 catégories)
Arts - Culture - Médias	Actu, médias	Actualités et médias
Emploi, formation	Culture et loisirs	Art et culture
Entreprises	Éducation, formation	Commerce et Économie
Finances - Bourse - Patrimoine	Espace B to B	Divertissement
Informatique - Internet	Forme et Santé	Enseignement et Formation
Infos - Météo	Mes Courses	Exploration géographique
Jeux - Consoles	Nature et sciences	Informatique et Internet
Loisirs - Passions	Nouvelles technologies	Institutions et politique
Santé	Société, Vie pratique	Références et annuaires
Savoir - Éducation	Sorties, spectacles	Santé
Sciences - Techniques	Sport et détente	Sciences et technologies
Shopping	Voyage, géographie	Sciences humaines
Sports		Société
Vie quotidienne - Société		Sports et loisirs
Voyages - Tourisme		

Au terme de cette utilisation combinée des annuaires, de l'identification des services de recherche (moteurs, annuaires), de communication (WebMail, WebChat, forums) et de ceux fournis par les portails généralistes par *CatService*, et du marquage des sites pornographiques, les taux de couverture des sessions par les descriptifs sont sensiblement améliorés. Globalement, les sessions sont décrites à 48 % en termes de durée ; la part des sessions couvertes à plus de 50 % – celles que l'on pourra exploiter – augmente nettement, et oscille entre 46 et 53 % en fonction du couple annuaire-données retenu (voir Tableau 5.32).

Tableau 5.32. Part de l'ensemble des sessions couvertes à plus de 50 % pour chaque couple annuaire-données (annuaires, CatService, catégorie « porno »)

	BibUsages	SN00-02	SN2002
MSN	49,0 %	46,4 %	52,3 %
Nomade	48,4 %	48,7 %	52,4 %
Yahoo	46,3 %	49,0 %	53,5 %

Parmi ces sessions bien couvertes, on compte généralement entre 30 et 35 % de sessions complètement décrites (environ 15 % de l'ensemble des sessions). Pour les autres, le taux de couverture est également réparti entre 50 et 100 %.

> ***Synthèse.*** *L'utilisation des annuaires seuls pour décrire les contenus visités dans les parcours se heurte à un double problème : d'une part, seules 25 à 30 % des sessions sont suffisamment décrites pour être exploitées ; d'autre part, les pages vues sur les portails généralistes sont rattachées à une catégorie unique dans les annuaires, ce qui est réducteur par rapport à la diversité de leur offre, et très gênant du fait de leur forte audience. La mobilisation des descriptions issues de* CatService *permet de lever ce double biais, grâce à une description fine des différents services sur les portails ; l'adjonction à ce dispositif d'une catégorie « pornographie » permet finalement de décrire correctement la moitié des sessions des panels, ce qui constitue une base solide pour l'analyse des parcours.*

5.3.2 Contenus visités

En travaillant sur les sessions décrites par les annuaires et *CatService* pour les portails généralistes pour plus de la moitié de leur durée, nous avons pour chaque session entre trente et trente-trois descripteurs de contenu, et pouvons commencer à examiner le contenu et la diversité des sessions en termes de thématiques et de services.

Panorama des types de contenus visités

En matière de diversité des contenus visités, le nombre de catégories différentes décrivant chaque session nous fournit un bon indice de l'éclatement des sessions : pour chaque jeu de données et chaque annuaire, nous avons examiné le nombre de descripteurs distincts pour chaque session. En premier lieu, on remarque la grande homogénéité des neuf résultats obtenus : quels que soient les données et l'annuaire retenus, les chiffres relatifs au nombre de catégories sont très similaires, ce qui tend à montrer que les comportements sont assez semblables et généraux en ce qui concerne l'éclatement thématique / fonctionnel de chaque session.

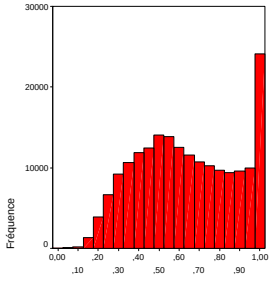
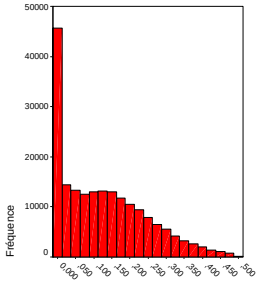
Les résultats eux-mêmes montrent une faible dispersion, avec en moyenne 3,5 catégories différentes visitées, et une médiane à 3 (voir Tableau 5.33 pour le calcul sur les données SN2002). D'autant plus que l'utilisation des descriptifs de *CatService* fait entrer dans ce compte les pages d'accueil des fournisseurs d'accès, points de passage rapides et non significatifs dans la session.

Tableau 5.33. SN2002, nombre de catégories descriptives par session

		MSN	Nomade	Yahoo
Moyenne		3,5	3,5	3,4
Médiane		3	3	3
Quintiles	20	1	1	1
	40	2	2	2
	60	4	4	4
	80	5	5	5

Dans la majorité des cas, la catégorie qui occupe le plus de temps dans la session représente une part importante de la durée totale de la session. Ici encore, les résultats sont très similaires entre les trois jeux de données et les trois annuaires (voir Tableau 5.34 pour le calcul sur les données SN2002) : dans l'ensemble, la catégorie la plus vue occupe presque les deux tiers de la durée de la session en moyenne, et ce quelle que soit la catégorie, tandis que la deuxième catégorie occupe un peu plus de 10 % de la session seulement. La troisième catégorie a un taux de couverture plus faible encore, proche de 3,5 % dans la plupart des cas.

Tableau 5.34. Part des catégories 1 et 2 dans la durée des sessions bien décrites –SN2002

	Catégorie la plus vue			Catégorie n° 2		
	MSN	Nomade	Yahoo	MSN	Nomade	Yahoo
Moyenne	63,7 %	63,8 %	64,1 %	12,7 %	12,6 %	12,3 %
Médiane	62,2 %	62,4 %	62,8 %	10,7 %	10,6 %	10,1 %
Distribution des valeurs						

Sur cette base, une première approche des contenus des sessions consiste à examiner la catégorie descriptive qui occupe la durée la plus importante dans la session. Nous pratiquons en premier lieu cet examen sur les données SensNet en 2002, les plus représentatives des usages généraux des internautes.

Tableau 5.35. Répartition des sessions par catégorie la plus forte dans la session, MSN et Nomade – SN2002[1]

MSN et CatService	%
TS - WebMail	16,6 %
PG - Page Accueil	12,2 %
Informatique - Internet	9,9 %
Finances - Bourse - Patrimoine	5,6 %
TS - WebChat	5,2 %
Entreprises	4,7 %
Jeux - Consoles	4,5 %
PORNO	4,5 %
Infos - Météo	4,2 %
Arts - Culture - Médias	4,1 %
Loisirs - Passions	3,8 %
TS - Moteur	3,2 %
Vie quotidienne - Société	3,1 %
PG - Informations	3,0 %
PG - Personnalisation	2,3 %
Shopping	2,1 %
PG - Non catégorisé	2,0 %
Sports	1,4 %
Savoir - Éducation	1,3 %
PG - Communication	1,3 %
Voyages - Tourisme	1,0 %

Nomade et CatService	%
TS - WebMail	16,5 %
PG - Page Accueil	12,1 %
Espace B to B	9,3 %
Société, Vie pratique	8,1 %
Sport et détente	7,9 %
Mes Courses	7,6 %
TS - WebChat	5,2 %
Actu, médias	4,5 %
PORNO	4,5 %
TS - Moteur	3,2 %
PG - Informations	3,0 %
Culture et loisirs	2,9 %
Nouvelles technologies	2,4 %
PG - Personnalisation	2,3 %
PG - Non catégorisé	2,0 %
Voyage, géographie	1,9 %
Éducation, formation	1,5 %
PG - Communication	1,3 %
Forme et Santé	0,8 %
PG - Divers	0,8 %
Nature et sciences	0,5 %

[1] Convention de nommage : les descripteurs issus de *CatService* sont préfixés par « PG » pour les portail généralistes, et « TS » pour les types de services (moteur, WebMail, WebChat, Forum).

Emploi, formation	0,9 %
PG - Divers	0,8 %
Santé	0,6 %
Sciences - Techniques	0,3 %
PG - Loisir En Ligne	0,3 %
PG - Annuaire	0,3 %
PG - Achat	0,3 %
PG - Page Perso	0,1 %
PG - Information Service	0,1 %
PG - Aide	0,1 %
PG - Information Produit	0,1 %
PG - Généralités	0,0 %
PG - Bourse	0,0 %

PG - Loisir En Ligne	0,3 %
PG - Annuaire	0,3 %
PG - Achat	0,3 %
Sorties, spectacles	0,2 %
PG - Page Perso	0,1 %
PG - Information Service	0,1 %
PG - Aide	0,1 %
PG - Information Produit	0,1 %
PG - Généralités	0,0 %
PG - Bourse	0,0 %

Clef de lecture : dans 16,6 % des sessions SN2002, c'est sur la catégorie « WebMail » que l'internaute passe le plus de temps dans la session.

Les résultats obtenus sur les trois bases descriptives sont relativement concordantes (voir Tableau 5.35 pour MSN et Nomade) : ils montrent en premier lieu l'importance des portails généralistes et des outils de communication dans les sessions Web, en particulier le WebMail. Ils attestent surtout la grande diversité des thématiques des parcours : les taux de représentation des catégories d'annuaires sont assez équilibrés, et correspondent globalement au nombre de sites présentés par l'annuaire dans la catégorie correspondante.

En revanche, on se gardera bien de donner à partir de ces chiffres des conclusions sur les usages d'Internet en général : les sessions sont ici considérées globalement, indépendamment de l'utilisateur, et rien ne nous renseigne ici sur le nombre de panélistes concernés. Par exemple, la catégorie « Finance – Bourse – Patrimoine » de MSN est la plus vue dans 5,6 % des sessions, mais ce comportement touche plus d'un tiers des utilisateurs du panel.

Nous verrons par la suite dans quelle mesure les différents services et thèmes des parcours sont répartis chez les utilisateurs. Nous avons déjà pu observer dans les pratiques générales d'Internet de grandes disparités entre internautes, tant dans l'intensité des pratiques que dans les différents protocoles et services utilisés, et l'on imagine bien que tout le monde ne s'intéresse pas à tout. Pour l'heure, on se contentera de remarquer, au niveau des sessions, que la diversité est de mise et que les pratiques reflètent la diversité de l'offre de contenus et de services.

Homogénéité à l'intérieur des sessions ?

On conclurait volontiers, sur la base de ces chiffres, que les sessions Web sont globalement mono-thématiques ; ce serait sans compter sur la diversité que masquent ces taux de couverture calculés sur l'ensemble des sessions. Comme nous avons déjà pu le remarquer, les sessions sont très différentes quant au nombre de sites différents visités : dans la mesure où les annuaires indexent principalement des sites plus que des pages, les descriptions de contenu sont directement influencées par le nombre de sites vus dans la session.

Ces disparités ont une première influence sur les taux de couverture des sessions par les descriptifs. Si l'on discrétise le nombre de sessions en quatre modalités aux effectifs similaires, les taux de couverture sont globalement semblables et oscillent

entre 42 et 44 % en durée pour SN2002 (résultats similaires pour les autres jeux de données) ; la part des sessions « bien » décrites est elle aussi stable, entre 39 et 44 %. Par contre, la distribution du taux de couverture est très variable (voir Figure 5.11) : lorsqu'un ou deux sites sont vus, dans la grande majorité des cas, la session est soit complètement décrite, soit pas du tout. À mesure qu'augmente le nombre de sites différents visités dans la session, la description quitte ce comportement binaire : au-delà de cinq sites différents visités dans la session, on a bien peu de chances d'avoir une description complète du parcours de l'internaute.

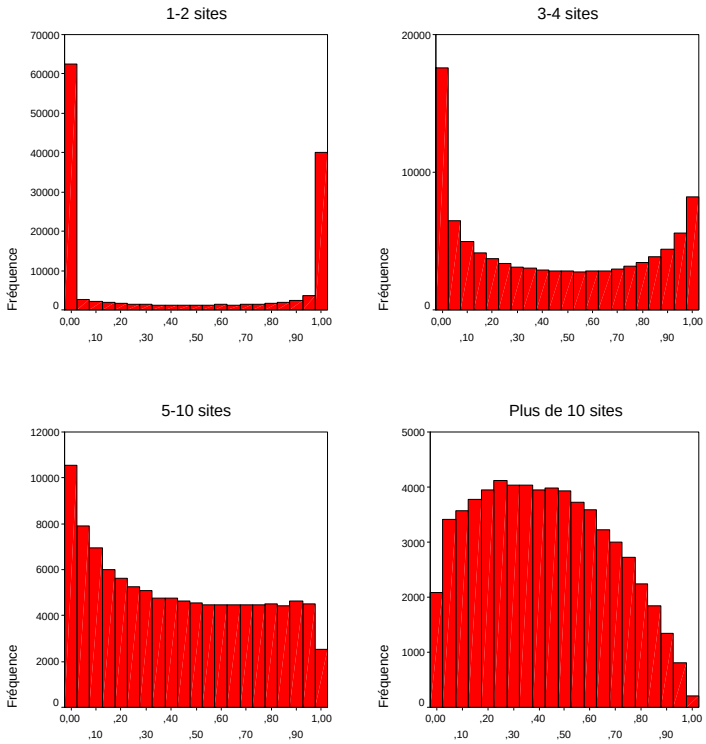


Figure 5.11. Nombre de sites distincts et taux de couverture par les descriptifs de sessions – SN2002, Yahoo-CatService-Porno

Ces éléments affectent directement le nombre de descriptifs différents représentés dans les sessions (voir Tableau 5.36). Malgré le recours à *CatService* qui descend à l'échelle des services pour les portails généralistes et multiplie le nombre possible de catégories descriptives pour ce type de sites, une session impliquant la visite de moins de quatre sites ne relève que de deux ou trois descriptifs en moyenne, contre six lorsque plus de dix sites différents sont vus.

Tableau 5.36. Nombre de sites et moyenne / médiane du nombre de catégories descriptives par session – SN2002

	MSN	Nomade	Yahoo
1-2 site	1,99 / 2	2,00 / 2	1,99 / 2
3-4 sites	3,00 / 3	2,99 / 3	2,89 / 3
5-10 sites	4,22 / 4	4,17 / 4	3,99 / 4
Plus de 10 sites	6,05 / 6	6,00 / 6	5,63 / 5

Corrélativement, la catégorie descriptive la plus représentée dans les sessions occupe de moins en moins de place à mesure que le nombre de sites et de descripteurs augmente (voir Tableau 5.37).

Tableau 5.37. Part de la catégorie la plus vue dans la durée des sessions bien décrites – SN2002

	MSN	Nomade	Yahoo
1-2 site	82 %	82 %	82 %
3-4 sites	64 %	64 %	65 %
5-10 sites	53 %	53 %	54 %
Plus de 10 sites	44 %	44 %	45 %

De 82 % de la durée totale de la session en moyenne lorsqu'un ou deux sites sont vus, elle n'en représente que 44 % lorsque le nombre de sites est supérieur à dix. La part de la deuxième catégorie la plus représentée dans les sessions reste quant à elle relativement stable : elle représente en moyenne 10 % de la durée des sessions lorsque peu de sites sont vus, et 13 % pour les plus riches. L'allongement des sessions s'accompagne donc d'une diversification de leurs contenus. Si l'homogénéité est de mise pour les sessions courtes, où peu de sites sont vus, elle est mise à mal dès lors que plus de cinq sites différents sont vus. Ce constat nous interdit de considérer les contenus visités de manière globale, et invite à les examiner séparément pour chaque groupe de sessions en fonction du nombre de sites vus dans la session.

Tableau 5.38. Répartition des sessions bien décrites par catégorie la plus vue dans la session – SN2002, 1-2 sites distincts visités

MSN	**%**
PG - Page Accueil	22,9
TS - WebMail	17,5
Informatique - Internet	13,1
Finances - Bourse	5,5
PG - Informations	4,1
Jeux - Consoles	3,7
Entreprises	3,7
PG - Personnalisation	3,5
Infos - Météo	3
Arts - Culture - Médias	2,7
Autres	20,3

Nomade	**%**
PG - Page Accueil	23,1
TS - WebMail	17,7
Espace B to B	13
Société, Vie pratique	7,2
Sport et détente	5,4
Mes Courses	5,3
PG - Informations	4,2
PG - Personnalisation	3,5
Actu, médias	3
TS - WebChat	2,5
Autres	15,3

Yahoo	**%**
PG - Page Accueil	24
Commerce et Economie	22,5
TS - WebMail	18,3
PG - Informations	4,3
PG - Personnalisation	3,6
Sports et loisirs	3,5
Exploration géographique	2,8
TS - WebChat	2,5
PG - Non catégorisé	2,5
Actualités et médias	2,4
Autres	13,6

Clef de lecture : si l'on décrit les sessions par les catégories MSN et *CatService*, la catégorie de *CatService* « WebMail » est la plus représentée dans 17,5 % des sessions.

Deux exemples illustreront le bien-fondé de cette démarche : si l'on compare la catégorie la plus représentée dans les sessions d'un ou deux sites d'une part, et de plus de dix sites d'autre part, les résultats sont sensiblement différents. Dans le premier cas, la page d'accueil des portails généralistes domine, suivie du WebMail (voir Tableau 5.38 ci-dessus). On peut faire l'hypothèse, pour ces sessions, d'une visite rapide et très ciblée de services d'information ou de communication, l'utilisateur sachant très bien ce qu'il cherche.

Pour les sessions de plus de dix sites, au contraire, c'est la catégorie des sites pornographiques qui est la plus représentée dans 20 % des sessions (voir Tableau 5.39), contre 4,5 % toutes sessions confondues. Les moteurs de recherche entrent également en jeu, témoignant sans doute de comportements de recherche longue où l'utilisateur est amené à visiter beaucoup de pages de résultats sur des sites différents.

Tableau 5.39. Répartition des sessions bien décrites par catégorie la plus vue dans la session – SN2002, plus de 10 sites distincts visités

MSN	%
PORNO	18,9
TS - Moteur	8,1
Informatique - Internet	7,7
TS - WebMail	7
TS - WebChat	6,5
Arts - Culture - Médias	5,5
Finances - Bourse	5,5
Jeux - Consoles	5,2
Loisirs - Passions	4,4
Entreprises	4,4
Infos - Météo	4,1
PG - Page Accueil	3,5
Autres	19,2

Nomade	%
PORNO	18,2
Sport et détente	9,1
Mes Courses	8,9
TS - Moteur	7,9
Société, Vie pratique	7,5
TS - WebMail	6,6
Espace B to B	6,6
TS - WebChat	6,3
Actu, médias	5,3
Culture et loisirs	4,1
Nouvelles technologies	3,4
PG - Page Accueil	3,2
Autres	12,9

Yahoo	%
Commerce et Economie	25
PORNO	18
TS - Moteur	7,8
TS - WebMail	6,7
TS - WebChat	6,5
Sports et loisirs	4,8
Exploration géographique	4,8
Actualités et médias	3,6
PG - Page Accueil	3,3
Divertissement	2,7
Société	2,6
Art et culture	2,1
Autres	11

Ici encore, le panel BibUsages marque sa différence : pour ces sessions longues, les catégories MSN les plus présentes sont « Arts – Culture – Médias » (14,6 % des sessions), « TS – Moteur » (13,8 %) et « Loisirs – Passions » (12,7 %) ; les sites pornographiques ne sont pas en reste pour autant (11,1 %), mais sont en retrait par rapport au panel général des internautes à la même période.

Plus généralement, les différences de thèmes et de services au regard du nombre de sites visités et de la durée des sessions confirment la nécessité de subordonner l'étude des contenus à celle de la topologie des parcours. Forme, durée, rythme des sessions sont la trace de chaînes opératoires qui sont étroitement liées aux thèmes et aux services visés par l'utilisateur, et témoignent de contextes d'usage différenciés.

> ***Synthèse.*** *L'examen des contenus visités dans les sessions montre l'importance des portails généralistes et des outils de communication dans les usages du Web, en particulier le WebMail. Considérés de manière globale, ils attestent la grande diversité des thématiques des parcours, mais à l'échelle de la session, cette diversité est beaucoup plus restreinte. Si les sessions courtes sont fortement mono-thématiques ou mono-fonctionnelles, et renvoient à un cours d'action unique, l'allongement des sessions, en durée comme en nombre de sites, s'accompagne d'une diversification de*

leurs contenus. Elle est également corrélée aux types de contenus eux-mêmes : dans les sessions courtes, les services d'information ou de communication dominent, tandis que les sessions longues mettent en avant les sites pornographiques et les moteurs de recherche. Ces éléments attestent le lien fort entre topologie et contenu des parcours sur le Web.

5.4 Profils de sessions

De l'analyse des usages généraux d'Internet au détail de l'activité au sein des sessions, on a pu observer une grande diversité dans les comportements : l'intensité d'usage du Web variable selon les utilisateurs fait écho aux disparités dans la durée, l'éclatement, le rythme et les contenus de parcours. Une approche typologique des parcours doit permettre de rendre compte de cette diversité et des régularités qui s'y jouent.

5.4.1 Classification

Sur la base de l'observation précise des éléments rythmiques, temporels et topologiques des sessions, nous sommes maintenant en mesure de construire une classification des sessions. Il s'agit ici d'affiner les oppositions qui ont pu être mises à jour dans l'analyse des composantes topologiques prises séparément.

Variables retenues

L'examen des variables temporelles, du nombre de sites et des indicateurs nous permet de les sélectionner et de les organiser de manière pertinente. Nous avons en effet vu que les variables continues ici manipulées masquent des distributions et des réalités qui nous obligent à les discrétiser : le nombre de sites différents vus dans une session ainsi que sa durée peuvent par exemple avoir des valeurs soit très faibles, soit extrêmes, mais il existe dans les faits une différence forte entre sessions très courtes avec peu de sites, et sessions plus longues et plus complexes.

De la même manière, le taux de linéarité, qu'il soit calculé à l'échelle du site ou de la page, peut avoir toutes les valeurs comprises entre 0 et 1, mais la valeur 1 correspond à une session linéaire, classe à part qui implique mécaniquement certaines valeurs pour les autres taux (taux de concentration, nombre de *back*, etc.). Dans tous les cas, les chiffres figurent une continuité là où, dans les pratiques, on observe des réalités et des comportements bien distincts.

Partant, nous avons discrétisé l'ensemble des variables pertinentes pour l'analyse en 3 à 5 classes de manière à obtenir des groupes homogènes en termes d'effectifs et surtout en termes de comportements sous-jacents ; le Tableau 5.40 présente la liste de ces variables et des différentes modalités construites pour chacune.

Tableau 5.40. Discrétisation des variables temporelles et topologiques retenues

Variables	Échelle d'analyse	Discrétisation	Notation	% des sess.
Nombre de sites distincts	-	1-2	1-2 sites	33,6 %
		3-4	3-4 sites	23,6 %
		5-10	5-10 sites	27,1 %
		>10	Plus de 10 sites	15,8 %
Durée de la session	-	1-3	1-3 min.	24,6 %
		4-13	4-13 min.	25,0 %
		14-34	14-34 min.	24,1 %
		>35	Plus de 35 min.	26,3 %
Durée médiane par page	-	0-2	DPage – 0-2 sec.	22,2 %
		3-4	DPage – 3-4 sec.	12,7 %
		5-9	DPage – 5-9 sec.	19,8 %
		10-15	DPage – 10-15 sec.	17,0 %
		>16	DPage – 16 sec. et plus	28,4 %
Durée médiane par site	-	0-9	DSite – 0-9 sec.	7,6 %
		10-19	DSite – 10-19 sec.	11,5 %
		20-29	DSite – 20-29 sec.	11,2 %
		30-1'09	DSite – 30 sec-1 min. 09	36,8 %
		>1'10	DSite – 1 min. 10 et plus	32,9 %
Taux de linéarité	Page	0-0,29	Pages – peu linéaire	2,8 %
		0,3-0,59	Pages – moy linéaire	24,6 %
		0,6-0,99	Pages – très linéaire	54,0 %
		1	Pages – linéaire	18,6 %
Taux de linéarité	Site	0-0,29	Sites – peu linéaire	13,0 %
		0,3-0,59	Sites – moy linéaire	25,9 %
		0,6-0,99	Sites – très linéaire	25,1 %
		1	Sites – linéaire	36,0 %
Taux de linéarité calculé sur la durée	Page	0-0,49	Pages – dur. Peu linéaire	35,5 %
		0,5-0,75	Pages – dur. Moy. linéaire	25,1 %
		0,76-0,99	Pages – dur. Très linéaire	20,6 %
			Pages – dur. Linéaire	18,8 %
Taux de linéarité calculé sur la durée	Site	0-0,29	Sites – dur. Peu linéaire	37,5 %
		0,3-0,59	Sites – dur. Moy. linéaire	13,8 %
		0,6-0,99	Sites – dur. Très linéaire	12,7 %
		1	Sites – dur. Linéaire	36,0 %
Concentration des revisites	Page	1	P – Conc. nulle	20,6 %
		1-1,9	P – Conc. faible	30,1 %
		2-2,9	P – Conc. moyenne	17,7 %
		3 et plus	P – Conc. forte	13,1 %
		-	P – Conc. undef (pas de revisite)	18,6 %
Concentration des revisites	Site	1	S – Conc. nulle	18,1 %
		1-1,9	S – Conc. faible	10,9 %
		2-3,49	S – Conc. moyenne	18,4 %
		3,5 et plus	S – Conc. forte	16,5 %
		-	S – Conc. undef (pas de revisite)	36,0 %
Nombre d'actions back	Page	0	P back – aucune	31,3 %
		1-3	P back – moyen	33,1 %
		4 et plus	P back – beaucoup	35,6 %

Nombre d'actions back	Site	0 1-4 5 et plus	S back – aucune S back – moyen S back – forte	39,7 % 36,4 % 23,9 %

Classification sur les données SN2002

Nous avons ainsi douze variables, pour un total de cinquante modalités distinctes. Afin d'avoir une vue la plus représentative possible des différents types de sessions que l'on peut rencontrer dans les pratiques, nous avons pratiqué préférentiellement une classification sur les données SN2002. Celle-ci servira de base de référence ensuite pour l'analyse des sessions observées dans les données SN00-02 et BibUsages.

En premier lieu, les résultats de l'analyse en composantes multiples pratiquée sur les 400 000 sessions de SensNet en 2002 sont encourageants (voir Tableau 5.41) : les dix premiers axes factoriels représentent près de 60 % de la variance de l'échantillon, et les cinq premiers axes résument à eux seuls 44 % de l'information.

Tableau 5.41. Classification sur la topologie des sessions, SN2002 – valeurs propres

Numéro d'axe	Valeur propre	Pourcent	Pourcent cumulé
1	0,6028	19,0 %	19,0 %
2	0,3054	9,6 %	28,7 %
3	0,1898	6,0 %	34,7 %
4	0,1518	4,8 %	39,5 %
5	0,1321	4,2 %	43,6 %
6	0,1120	3,5 %	47,2 %
7	0,1040	3,3 %	50,5 %
8	0,0981	3,1 %	53,6 %
9	0,0940	3,0 %	56,5 %
10	0,0908	2,9 %	59,4 %

On est donc bien fondé à effectuer une classification ascendante hiérarchique sur les résultats de l'ACM, en prenant en compte les dix premiers axes factoriels. Trois coupures de l'arbre sont proposées à l'issue de la classification, partitionnant l'échantillon en cinq, huit ou dix classes. Nous avons retenu la partition en cinq classes, qui présente le saut le plus important et les variabilités intra et inter-classes les plus significatives.

Les résultats de cette classification donnent des groupes bien distincts. La Figure 5.12 p. 219 représente la projection des individus sur les axes 1 et 2 (29 % de l'information représentée) : elle oppose très nettement les classes 1 et 2 aux trois autres. La projection des modalités des variables sur le graphique montre une opposition forte sur l'axe 1 entre sessions linéaires et non linéaires, tandis que l'axe 2 semble distinguer, au sein des sessions non linéaires, celles qui le sont peu et celles qui le sont beaucoup.

Dans une vue basée sur les axes 3 et 4 (Figure 5.13, p. 220), on retrouve cette homogénéité des classes 3, 4 et 5 qui forment un noyau central, autour duquel s'opposent les classes 1 et 2. Ici, sont essentiellement distinguées la linéarité inter-sites ainsi que le taux de concentration inter-site (axe 3), et le temps médian passé par visite de site (axe 4).

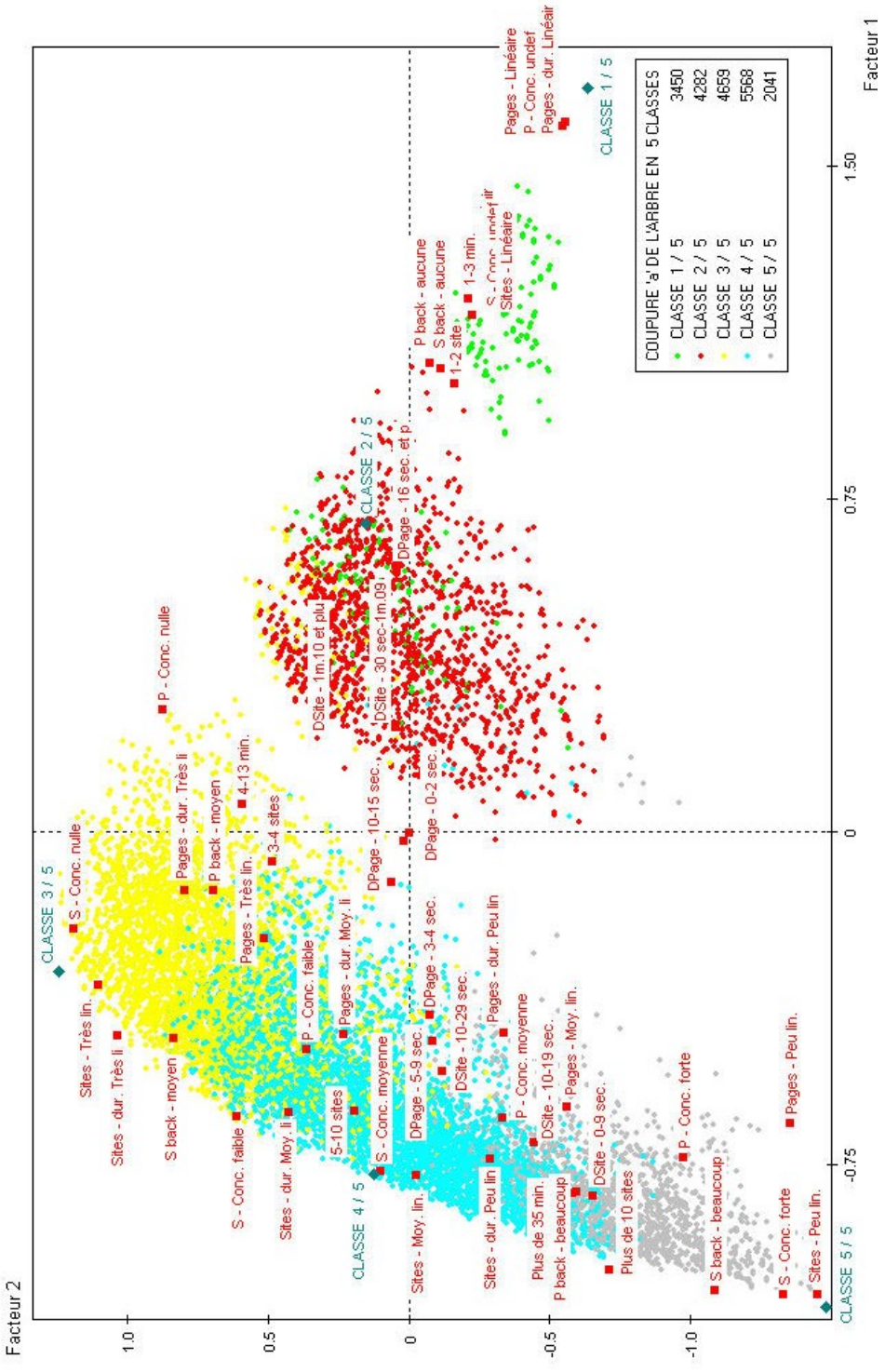


Figure 5.12. Classification sur les indicateurs topologiques, SN2002 - axes 1 et 2

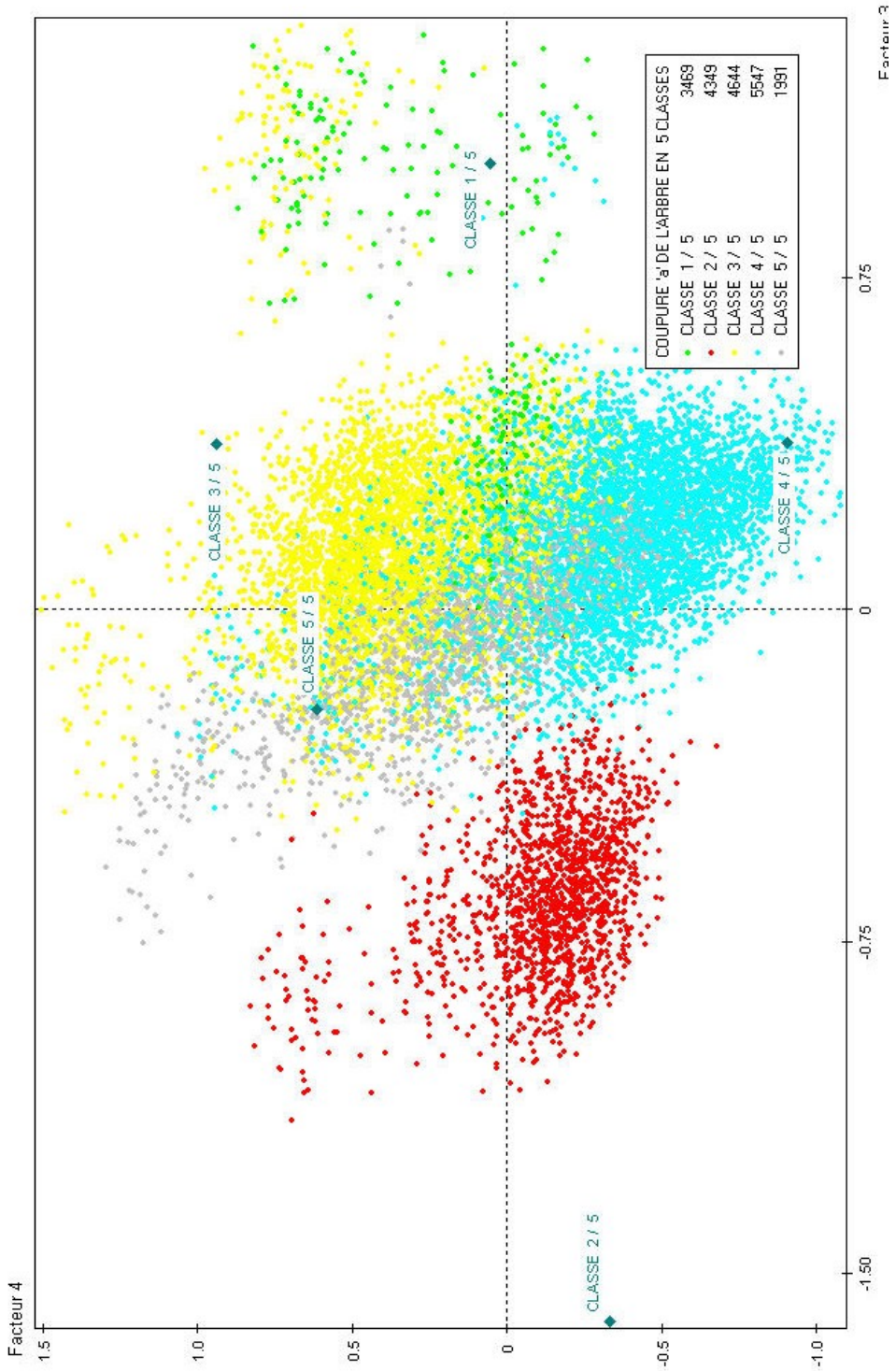


Figure 5.13. Classification sur les indicateurs topologiques, SN2002 - axes 3 et 4

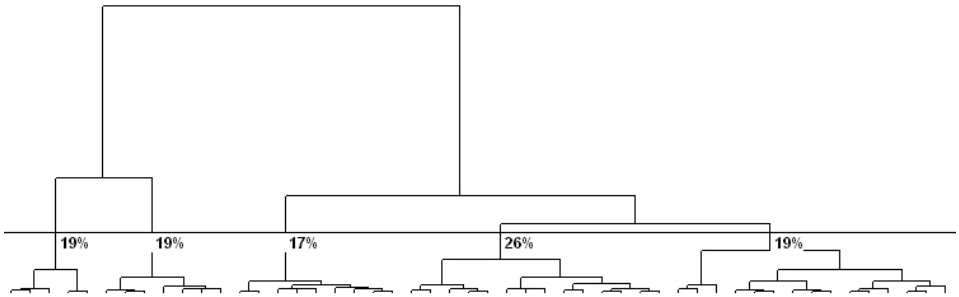


Figure 5.14. Classification sur la topologie des sessions, SN2002 – dendrogramme

Les classes 1 et 2 forment ainsi un groupe homogène de sessions qui s'opposent aux trois autres groupes, ce qu'illustre le dendrogramme représentant la classification (voir Figure 5.14). Les sessions de ce groupe sont courtes, linéaires ou quasi-linéaires, peu de sites sont vus et l'utilisateur passe plutôt du temps sur chaque site visité ; on semble être ici dans le contexte de visites ciblées où l'internaute accomplit des actions bien précises, qui ne l'amènent pas à sortir de son champ d'action.

Le deuxième grand groupe de sessions comprend trois classes distinctes, qui sont globalement caractérisées par une linéarité moyenne à faible, un nombre plus élevé de sites et de pages, et un allongement dans la durée. Le corpus voit ainsi s'opposer des sessions plutôt courtes et directes aux sessions allongées et plus complexes.

> ***Synthèse.*** *La structure particulière des indicateurs topologiques nous oblige à y opérer une discrétisation manuelle tenant compte de la spécificité des comportement qu'ils représentent. L'analyse en composantes multiples et la classification des sessions menées sur cette base traduisent la nécessité de subordonner l'analyse des contenus à celle des modes d'activité et des structures de navigation. Cinq groupes sont distingués, soumis à une opposition globale entre sessions courtes et linéaires et sessions allongées et complexes.*

5.4.2 Profils de sessions

Groupe des sessions courtes et directes

La première classe (17,4 % des sessions) est constituée de sessions courtes (1 à 3 minutes) et linéaires, et agrège des variables qui se retrouvent étroitement liées pour certaines valeurs particulières (voir Tableau 5.42 ci-dessous) : linéarité à l'échelle de la page et du site, absence de mouvements de *back*, faible nombre de sites visités (un ou deux).

On nommera cette classe « parcours éclairs » : avec des durées courtes (une à trois minutes, contre 35 minutes en moyenne pour l'ensemble des sessions), deux sites au plus, moins d'une minute par site, ces sessions Web sont fondamentalement caractérisées par leur brièveté. La Figure 5.15 en donne deux illustrations parmi les individus les plus représentatifs de la classe.

Tableau 5.42. SN2002, classification sur les indicateurs topologiques – Parcours éclairs

% dans la classe	% global	Variable	Échelle d'analyse	Modalités caractéristiques
100 %	18,6 %	Linéarité	Page	Linéaire
100 %	18,6 %	Concentration des revisites	Page	Conc. undef
100 %	18,8 %	Linéarité sur la durée	Page	Linéaire
100 %	31,3 %	Nb action back	Page	Aucune
92,4 %	36,0 %	Linéarité sur la durée	Site	Linéaire
92,4 %	36,0 %	Linéarité	Site	Linéaire
92,4 %	36,0 %	Concentration	Site	Conc. undef
79,5 %	24,6 %	Durée de session	-	1-3 min.
93,4 %	39,7 %	Nb actions back	Site	Aucune
81,9 %	33,6 %	Nb sites distincts	-	1-2 site
61,0 %	28,4 %	Durée médiane par visite	Page	16 sec. et plus
60,4 %	36,8 %	Durée médiane par visite	Site	30 sec-1m.09

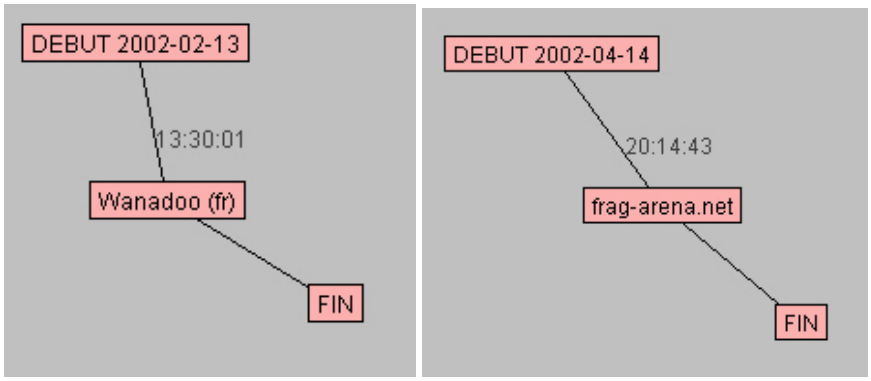


Figure 5.15. SN2002 - exemples typiques de la classe « parcours éclairs » (inter-site)

La classe 2 des « parcours ciblés » (19,9 % de l'ensemble) se rapproche de la classe 1 par le faible nombre de sites (1 à 2 sites) et une forte présence de sessions linéaires en inter-sites (voir Tableau 5.43).

Tableau 5.43. SN2002, classification sur les indicateurs topologiques – Parcours ciblés

% dans la classe	% global	Variable	Échelle d'analyse	Modalités caractéristiques
100 %	36,0 %	Concentration	Site	Conc. undef
100 %	36,0 %	Linéarité	Site	Linéaire
100 %	36,0 %	Linéarité sur la durée	Site	Linéaire
100 %	39,7 %	Nb actions back	Site	Aucune
71,9 %	32,9 %	Durée médiane par visite	Site	1m.10 et plus
69,5 %	33,6 %	Nombre de sites distincts	-	1-2 site
49,7 %	20,6 %	Concentration	Page	Conc. nulle
55,5 %	33,1 %	Nb actions back	Page	Moyen
73,5 %	54,0 %	Linéarité	Page	Quasi-linéaire

Par contre, à l'échelle de la page, les sessions ne sont pas linéaires, mais seulement « quasi-linéaires » (taux entre 0,6 et 1) : à l'intérieur d'un site, l'utilisateur est amené à revoir modérément certaines pages. Cette revisite semble due principalement aux mouvements de type *back*, dont le nombre est relativement important en regard de la durée générale des sessions du groupe. L'exemple de session donné ci-dessous illustre ces éléments : à l'échelle du site (Figure 5.16), la session est linéaire, avec deux sites visités ; à l'échelle de la page (Figure 5.17), la linéarité se trouve brisée par quelques « écarts » ponctuels.

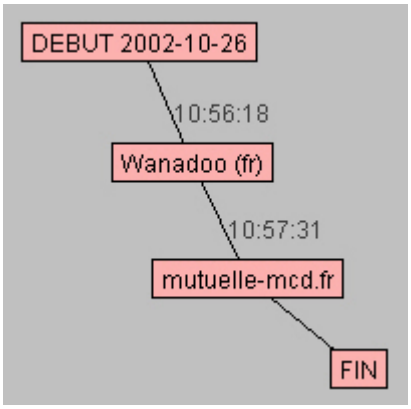


Figure 5.16. SN2002 - exemple typique de la classe « parcours ciblés »

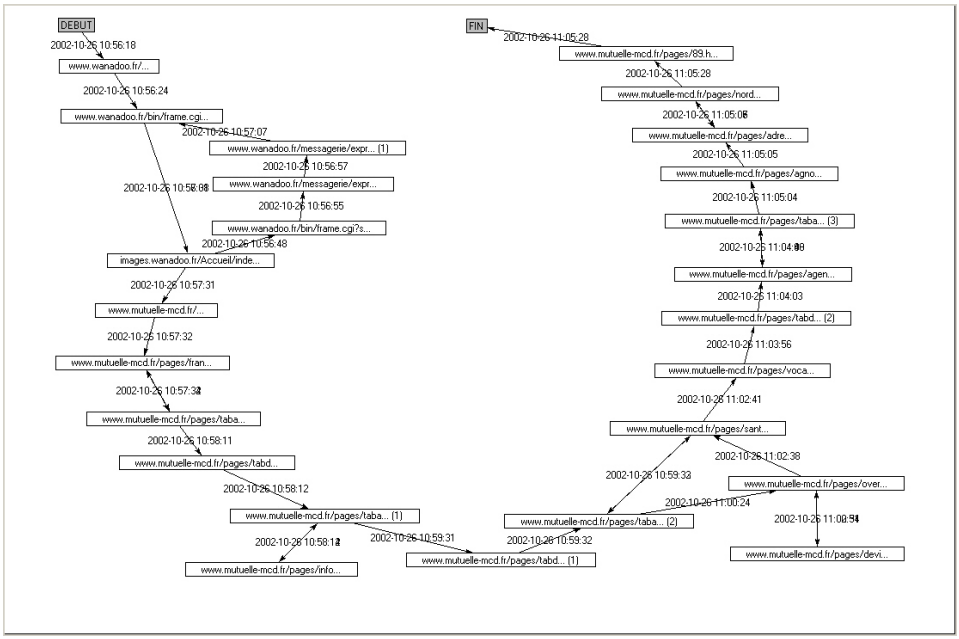


Figure 5.17. SN2002 - exemple de la classe « parcours ciblés » (inter-page)

Enfin, en termes de durée, ces sessions sont moins déterminées que celles du premier groupe, la durée n'intervenant pas comme modalité caractéristique de la

classe : il semble ici s'agir plutôt de « parcours ciblés », où les sites visés sont bien identifiés, mais la navigation au sein de ces sites s'allonge et se complexifie.

Sur le plan du contenu, parcours éclairs et parcours ciblés sont assez similaires : la projection des catégories d'annuaires et de *CatService* sur ces deux groupes montre une prédominance de la page d'accueil des portails généralistes pour le premier, et du WebMail pour le second, ce qui va de pair avec l'hypothèse de sessions très ciblées. À l'inverse, les services de recherche sur le Web sont complètement absents de ces sessions dont le faible nombre de sites implique qu'elles correspondent à des sites connus et mémorisés (page de démarrage, enregistrement dans des favoris) ou qu'il s'agisse de l'ouverture d'un lien à partir d'une source hors Web, comme un mail par exemple.

Groupe des sessions longues et complexes

La troisième classe, les « parcours à détours », regroupe 21,3 % des sessions et constitue la charnière entre les deux grands groupes : on a pu voir dans la Figure 5.12 (p. 219) que certains des individus de cette classe se confondent avec ceux de la classe 2. Ici, la linéarité à l'échelle du site n'est plus assurée, mais elle reste importante (voir Tableau 5.44), et la revisite de sites est principalement imputables à des mouvements de type *back*.

Tableau 5.44. SN2002, classification sur les indicateurs topologiques – Parcours à détours

% dans la classe	% global	Variable	Échelle d'analyse	Modalités caractéristiques
95,5 %	25,1 %	Linéarité	Site	Quasi-linéaire
78,0 %	18,1 %	Concentration	Site	Conc. nulle
85,1 %	36,4 %	Nb actions back	Site	Moyen
44,1 %	12,7 %	Linéarité sur la durée	Site	Quasi-linéaire
76,3 %	54,0 %	Linéarité	Page	Quasi-linéaire
36,8 %	20,6 %	Concentration	Page	Conc. nulle
36,8 %	20,6 %	Linéarité sur la durée	Page	Quasi-linéaire
49,8 %	33,1 %	Nb actions back	Page	Moyen
37,3 %	23,6 %	Nb sites distincts	-	3-4 sites
38,6 %	25,0 %	Durée de session	-	4-13 min.
39,2 %	27,1 %	Nb sites distincts	-	5-10 sites
23,0 %	13,8 %	Linéarité sur la durée	Site	Moyennement linéaire
31,3 %	24,1 %	Durée de session	-	14-34 min.

Ces mouvements de *back* ne sont pas articulés autour d'une page pivot, les taux de concentration restant nuls : on est ici dans le cas de parcours simples avec quelques digressions. La linéarité à l'échelle du site calculée sur la durée restant importante, ce qui indique que le temps passé sur des sites revus demeure faible. Dans ces sessions, on observe une véritable « colonne vertébrale » qui dirige le parcours, et les quelques détours ou boucles qui le jalonnent demeurent marginaux. Les exemples ci-dessous de sessions typiques de la classe le confirment (voir Figure

5.18) : dans cette classe des « parcours à détours », on observe des « pas de côté » et des boucles, mais aucune page ni site autour desquels s'articule la navigation.

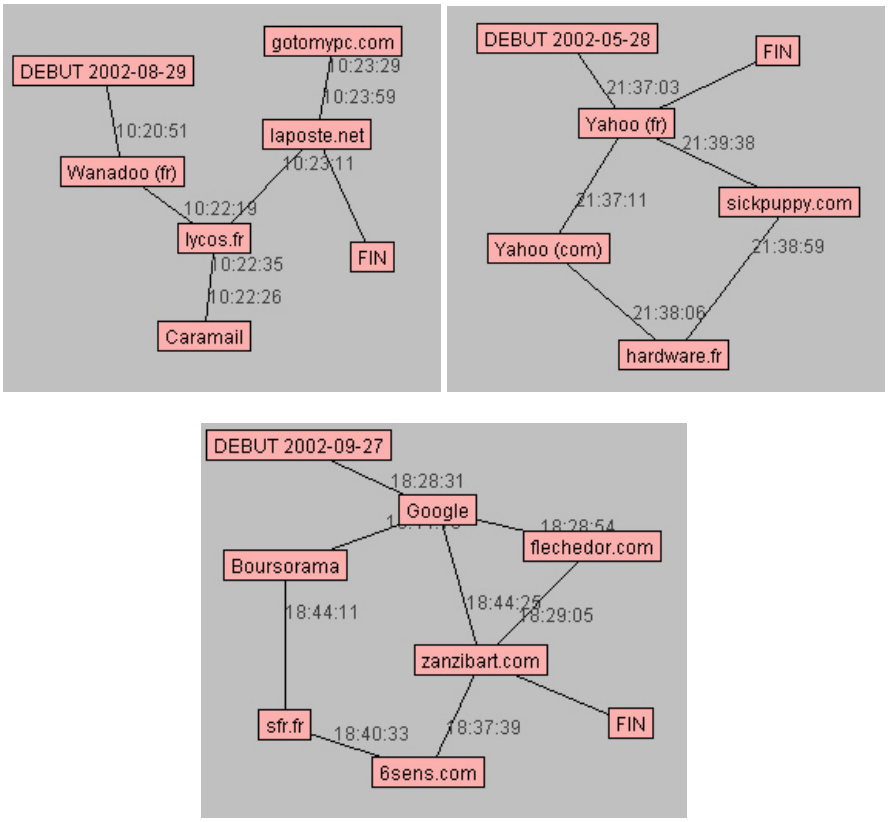


Figure 5.18. SN2002 - exemple typique de la classe « parcours à détours »

En termes de durée, cette classe est plutôt orientée vers des temporalités moyennes (moins d'un quart d'heure), mais avec une palette de sites différents visités pouvant aller jusqu'à dix. Ceci explique que les variables rythmiques (temps moyen par page et par site visité) n'entrent pas dans les variables les plus spécifiques de la classe.

Dans cette classe, on retrouve encore beaucoup de sessions où le WebMail occupe la première place en termes de durée, mais on observe une très nette diversification des thèmes des sessions. Les différentes catégories d'annuaires y sont représentées, avec un accent particulier pour celles relatives à la « vie pratique » : « Infos – météo » (MSN), « Société, vie pratique » (Nomade), « Exploration géographique » (catégorie de Yahoo qui renvoie plus particulièrement à des sites géolocalisés, correspondant à des services accessibles « hors Web » comme les associations, les institutions, etc.). Il semble que ces sessions soient plus orientées vers une pratique du type « pages jaunes » du Web : les moteurs de recherche y sont peu employés, ce qui montre que dans ce contexte, les internautes savent plutôt s'orienter et visitent des sites connus ou suivent des liens à partir de sites de confiance.

C'est dans la classe 4, que l'on va trouver une sur-représentation particulièrement forte des moteurs de recherche. Cette classe des « parcours à pivots » représente 26,4 % de l'ensemble des sessions. Les caractéristiques de la classe montrent un allongement et une complexification des sessions (voir Tableau 5.45) : la linéarité à l'échelle du site est moyenne, et certains sites-pivot apparaissent (concentration moyenne et faible au niveau du site, faible au niveau de la page). La faible linéarité à l'échelle du site en termes de durée montre que ces sites revisités occupent une part importante de la durée des sessions ; cette revisite peut se faire *via* les actions de type *back* (forte présence au niveau des pages comme des sites), mais également par des boucles.

Tableau 5.45. SN2002, classification sur les indicateurs topologiques – Parcours à pivots

% dans la classe	% global	Variable	Échelle d'analyse	Modalités caractéristiques
83,2 %	25,9 %	Linéarité	Site	Moyennement linéaire
57,9 %	18,4 %	Concentration	Site	Conc. moyenne
32,5 %	10,9 %	Concentration	Site	Conc. faible
62,8 %	36,4 %	Nb actions back	Site	Moyen
63,1 %	37,5 %	Linéarité	Site	Peu linéaire
52,7 %	30,1 %	Concentration	Page	Conc. faible
55,2 %	35,6 %	Nb actions back	Page	Beaucoup
30,7 %	15,8 %	Nb sites distincts	-	Plus de 10 sites
27,1 %	13,8 %	Concentration sur la durée	Site	Moyennement linéaire
41,5 %	26,3 %	Durée de la session	-	Plus de 35 min.
40,8 %	27,1 %	Nb sites distincts	-	5-10 sites
35,8 %	23,9 %	Nb actions back	Site	Beaucoup
37,1 %	25,1 %	Linéarité sur la durée	Page	Moyennement linéaire
65,4 %	54,0 %	Linéarité sur la durée	Page	Quasi-linéaire
33,5 %	24,1 %	Durée de la session	-	14-34 min.
25,4 %	17,7 %	Concentration	Page	Conc. moyenne
17,6 %	11,2 %	Durée médiane par visite	Site	10-29 sec.
32,4 %	24,6 %	Linéarité	Page	Moyennement linéaire
16,7 %	11,5 %	Durée médiane par visite	Site	20-19 sec.

Les exemples de sessions témoignent de cette diversité : dans la Figure 5.19, on observe une double boucle accolée, avec un seul mouvement de *back* à l'échelle du site. Les Figure 5.20 et Figure 5.21 montrent des mouvements plus complexes, mêlant boucles et retours-arrière, avec dans les deux cas deux sites qui servent de pivot à la visite des autres sites.

Dans ces retours et ces détours, l'utilisateur est amené à voir beaucoup de sites (plus de 10), et corrélativement la session s'allonge et dépasse la demi-heure. Cet allongement fait entrer dans les variables caractéristiques de la classe les éléments de rythme, absents de la classe des « parcours à détours » : la durée médiane par visite de site comprise entre 10 et 20 secondes – entre 5 et 10 secondes par visite de page – atteste une accélération de la navigation. Une part importante des sites et des pages ne semble être que traversée rapidement, ce qui rejoint l'utilisation forte de la fonction *back*.

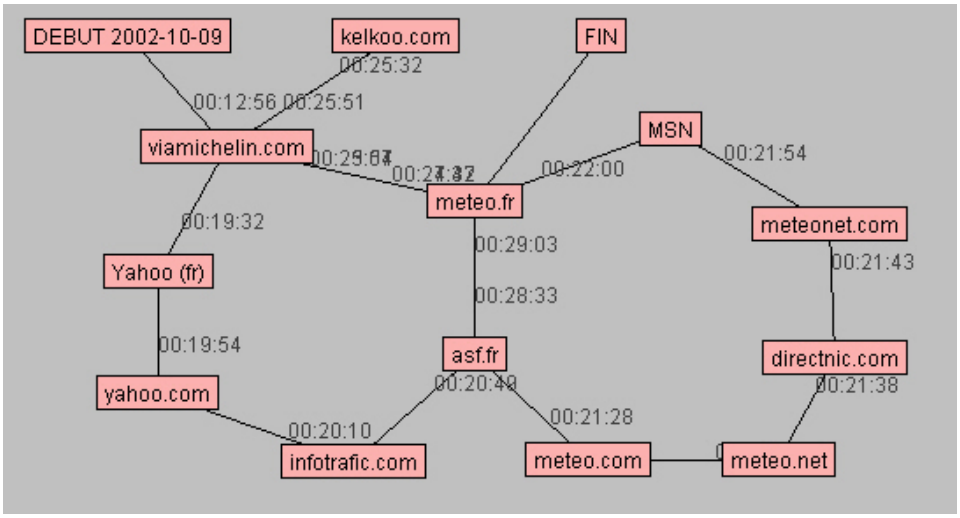


Figure 5.19. SN2002 - exemple typique de la classe « parcours à pivots »

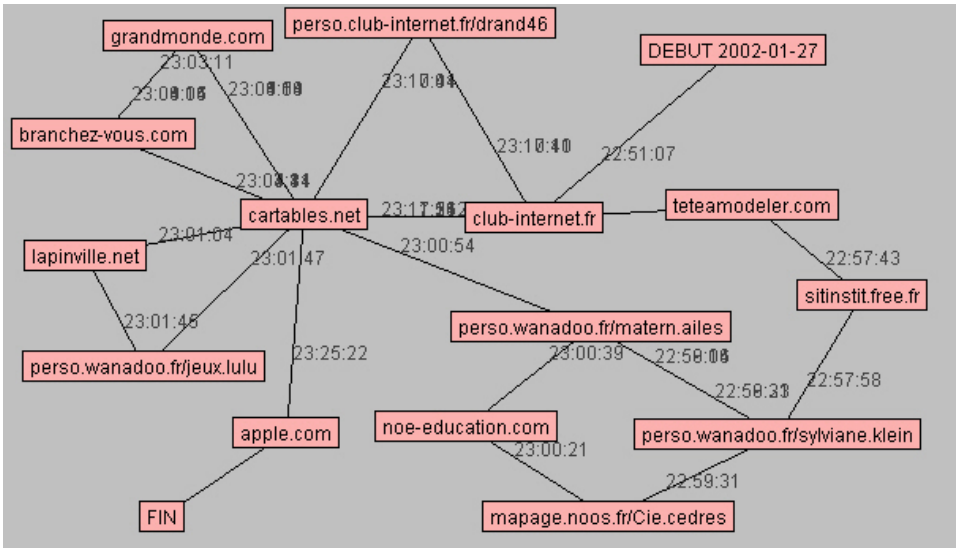


Figure 5.20. SN2002 - exemple typique de la classe « parcours à pivots »

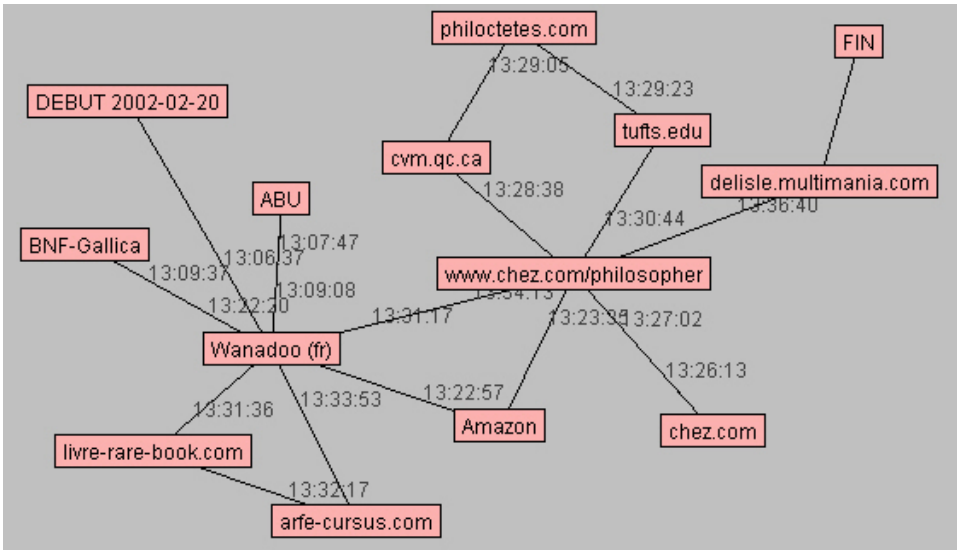


Figure 5.21. SN2002 – exemple typique de la classe « parcours à pivots »

Ces éléments morphologiques sont cohérents avec le recours important aux moteurs de recherche : l'internaute balaye des pages de résultats renvoyés par les moteurs, ou explore de nouveaux sites à partir de liens dans des pages-ressources. Sur le plan des thématiques de ces sessions, on observe un certaine hétérogénéité générale, mais une importante cohérence interne : chaque session semble axée autour d'un thème particulier autour duquel s'articule la recherche de l'internaute. Ce peut être la préparation d'un voyage comme on le voit dans l'exemple de la Figure 5.19, où l'internaute visite des sites d'information météorologique et de préparation d'itinéraire ; ou bien une recherche de textes philosophiques comme dans l'exemple de la Figure 5.21, qui amène l'utilisateur à visiter des bibliothèques numériques (Gallica, ABU), des sites de vente de biens culturels (Amazon, `livre-rare-book.com`) et des sites traitant de la philosophie. Sous-représentés dans les trois autres classes de parcours, les sites pornographiques sont ici très présents, ce qui rejoint le recours important aux moteurs de recherche sur lesquels on sait que les requêtes les plus fréquentes sont liées à ce thème.

La dernière classe, qui concerne 15 % des sessions, s'oppose autant aux « parcours éclairs » qu'aux « parcours à détours » : les caractéristiques principales de cette classe des « parcours éclatés » sont une forte concentration au niveau des sites, et une très faible linéarité des parcours (voir Tableau 5.46). On retrouve ici les sessions les plus longues des données, tant en nombre de sites qu'en durée. La fonction *back* des navigateurs est très utilisée tant au niveau des pages que des sites, et le temps moyen pour chaque passage sur un site est dans les moyennes basses.

Tableau 5.46. SN2002, classification sur les indicateurs topologiques – Parcours éclatés

% dans la classe	% global	Variable	Échelle d'analyse	Modalités caractéristiques
95,2 %	16,5 %	Concentration	Site	Conc. forte
81,6 %	13,0 %	Linéarité	Site	Peu linéaire
96,0 %	23,9 %	Nb actions back	Site	Beaucoup
86,4 %	37,5 %	Linéarité	Site	Peu linéaire
84,1 %	35,6 %	Nb actions back	Page	Beaucoup
63,7 %	26,3 %	Durée de la session	-	Plus de 35 min.
41,2 %	13,1 %	Concentration	Page	Conc. forte
40,1 %	15,8 %	Nb sites distincts	-	Plus de 10 sites
51,4 %	24,6 %	Linéarité	Page	Moyennement linéaire
25,1 %	7,6 %	Durée médiane par visite	Site	0-9 sec.
61,9 %	35,5 %	Linéarité	Page	Peu linéaire
29,0 %	11,5 %	Durée médiane par visite	Site	10-19 sec.
34,3 %	17,7 %	Concentration	Page	Conc. moyenne
9,3 %	2,8 %	Linéarité	Page	Peu linéaire
39,9 %	27,1 %	Nb sites distincts	-	5-10 sites
31,2 %	19,8 %	Durée médiane par visite	Page	5-9 sec.
19,1 %	11,2 %	Durée médiane par visite	Site	20-29 sec.
19,5 %	12,7 %	Durée médiane par visite	Page	3-4 sec.

Si l'on examine de plus près les sessions correspondantes, on constate que ces chiffres rendent compte de deux types de parcours distincts. D'une part, des sessions très éparpillées, où beaucoup de sites sont revus, et certaines boucles sont parcourues plusieurs fois, ce qu'illustrent les Figure 5.22 et Figure 5.23.

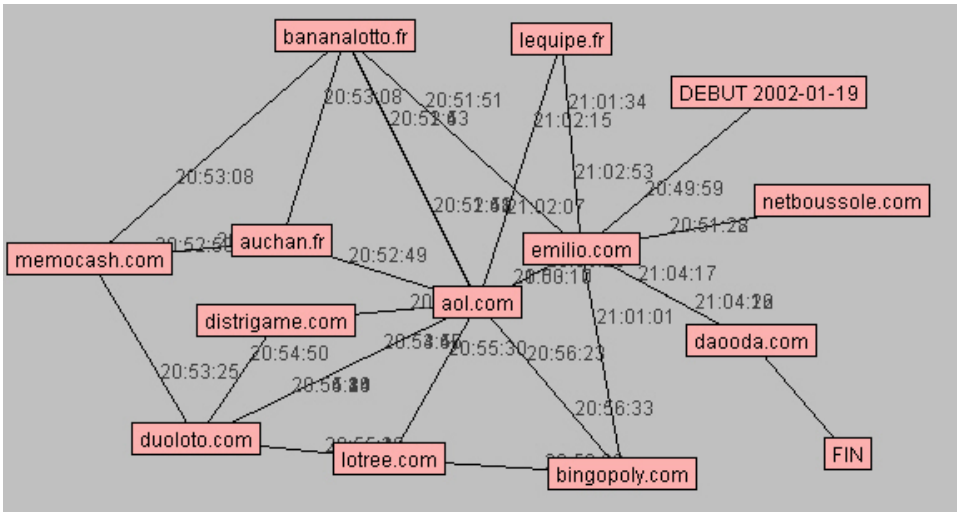


Figure 5.22. SN2002 – exemple typique de la classe « parcours éclatés »

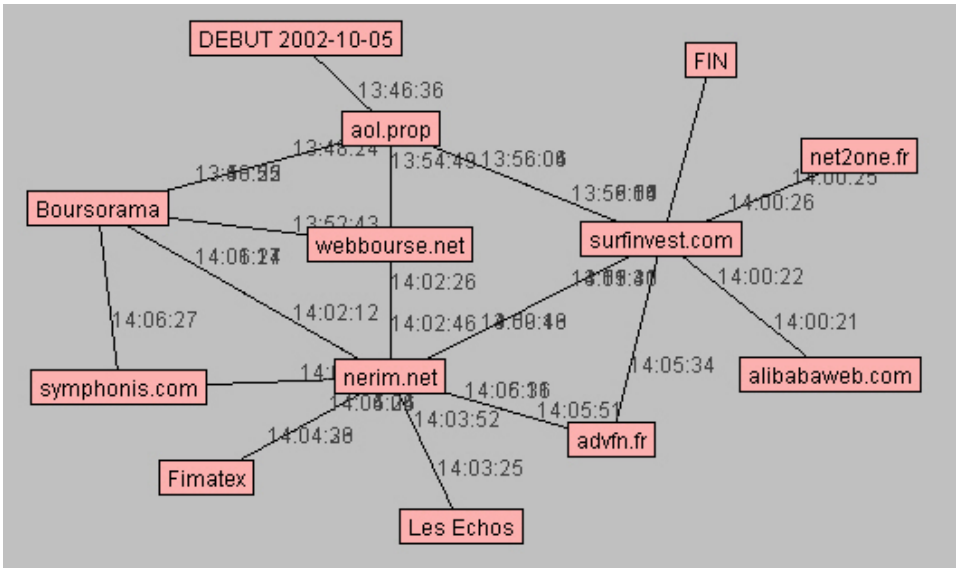


Figure 5.23. SN2002 – exemple typique de la classe « parcours éclatés »

D'autre part, des sessions où certains sites « complémentaires » ou certaines pages à taux de rafraîchissement élevé (les services de *chat*, par exemple) figurent dans les données une redondance et des échanges rapides entre deux sites ou deux pages. Ces éléments ont pour conséquence de figurer des taux de concentration très élevés, mais il s'agit dans la plupart des cas d'un biais des données. Ainsi, dans la session présentée Figure 5.24, les aller-retour entre les sites `locatorserver.net` et `cyberain.net` sont un artéfact par rapport au point de vue de l'utilisateur, `locatorserver.net` étant un site de bannières publicitaires mal identifié dans les données.

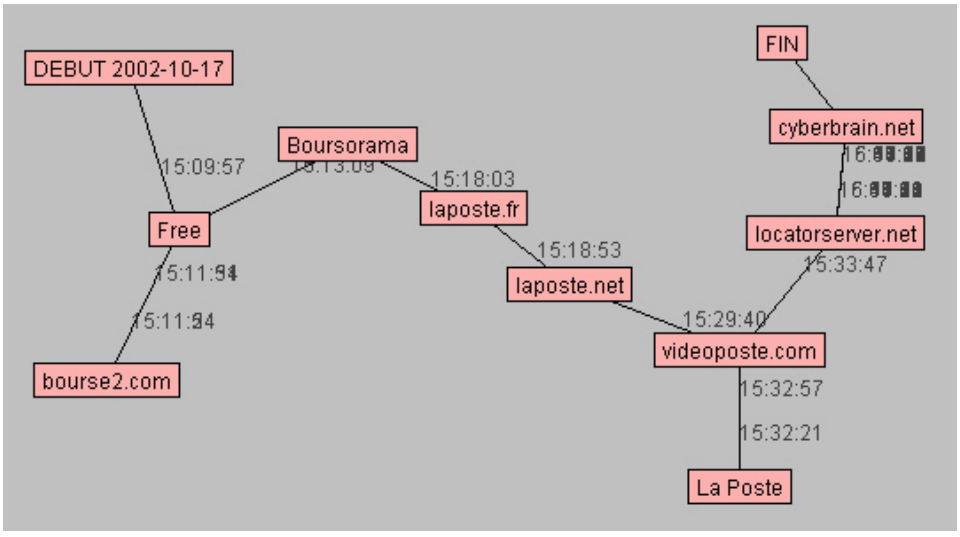


Figure 5.24. SN2002 – exemple typique de la classe « parcours éclatés »

Sur le plan des contenus visités, on retrouve dans ces « parcours éclatés » des contenus très diversifiés, et une diminution de la cohérence thématique globale : comme on a déjà pu l'observer en examinant les contenus des sessions, l'accroissement du nombre de sites visités s'accompagne d'une diversification des thématiques des sessions. En conséquence, aucun thème particulier ne ressort au sein de cette classe, sinon la pratique du WebChat qui implique de faibles taux de linéarité liés au rafraîchissement des pages et à l'allongement de la durée de sessions dans le cadre des échanges interpersonnels. Tout au plus observe-t-on une présence plus marquée des contenus orientés vers les activités ludiques (« Jeux – Consoles » dans MSN, « Sports et loisirs » pour Yahoo, « Sport et détente » chez Nomade, catégorie « pornographie ») d'une part et culturelles (« Culture et loisir » dans Nomade, « Arts – Culture » chez MSN) d'autre part, devançant de peu les activités orientées « vie pratique » impliquant des navigations plus longues, en particulier les catégories liées à la bourse et aux services bancaires.

Comparaison avec les autres jeux de données

Le même travail de classification pratiqué sur les deux autres jeux de données fournit globalement les mêmes résultats, et met en évidence les cinq groupes-types de navigation que nous venons de décrire (voir Figure 5.26 et Figure 5.27 ci-dessous). Ceci étant posé, pour comparer plus finement le positionnement de sessions de BibUsages et de SN00-02 par rapport au panel représentatif de 2002, nous devions travailler sur le même référentiel ; pour cela, nous avons inclus les sessions des panels BibUsages et SN00-02 en tant qu'individus illustratifs dans la classification pratiquée sur les données SN2002, et examiné leur position par rapport aux sessions du panel représentatif. Un premier élément de différenciation concerne la présence des différentes catégories de parcours dans les deux panels (voir Figure 5.25).

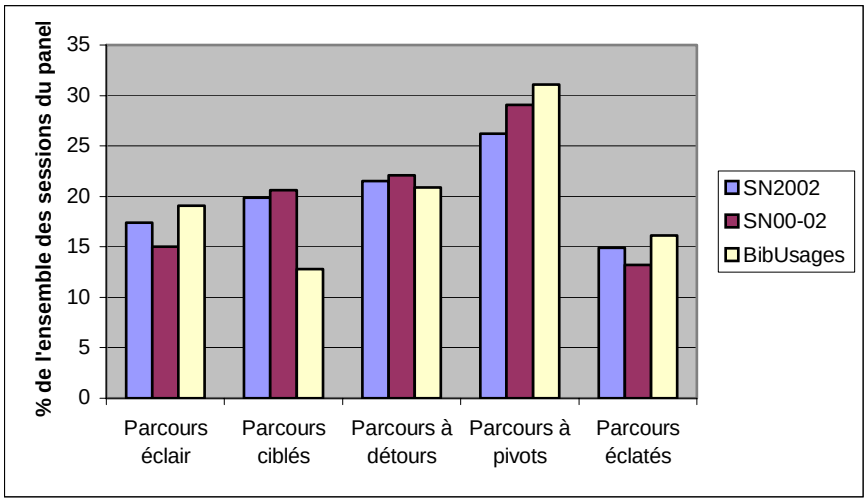


Figure 5.25. Répartition par groupe de sessions pour chaque jeu de données

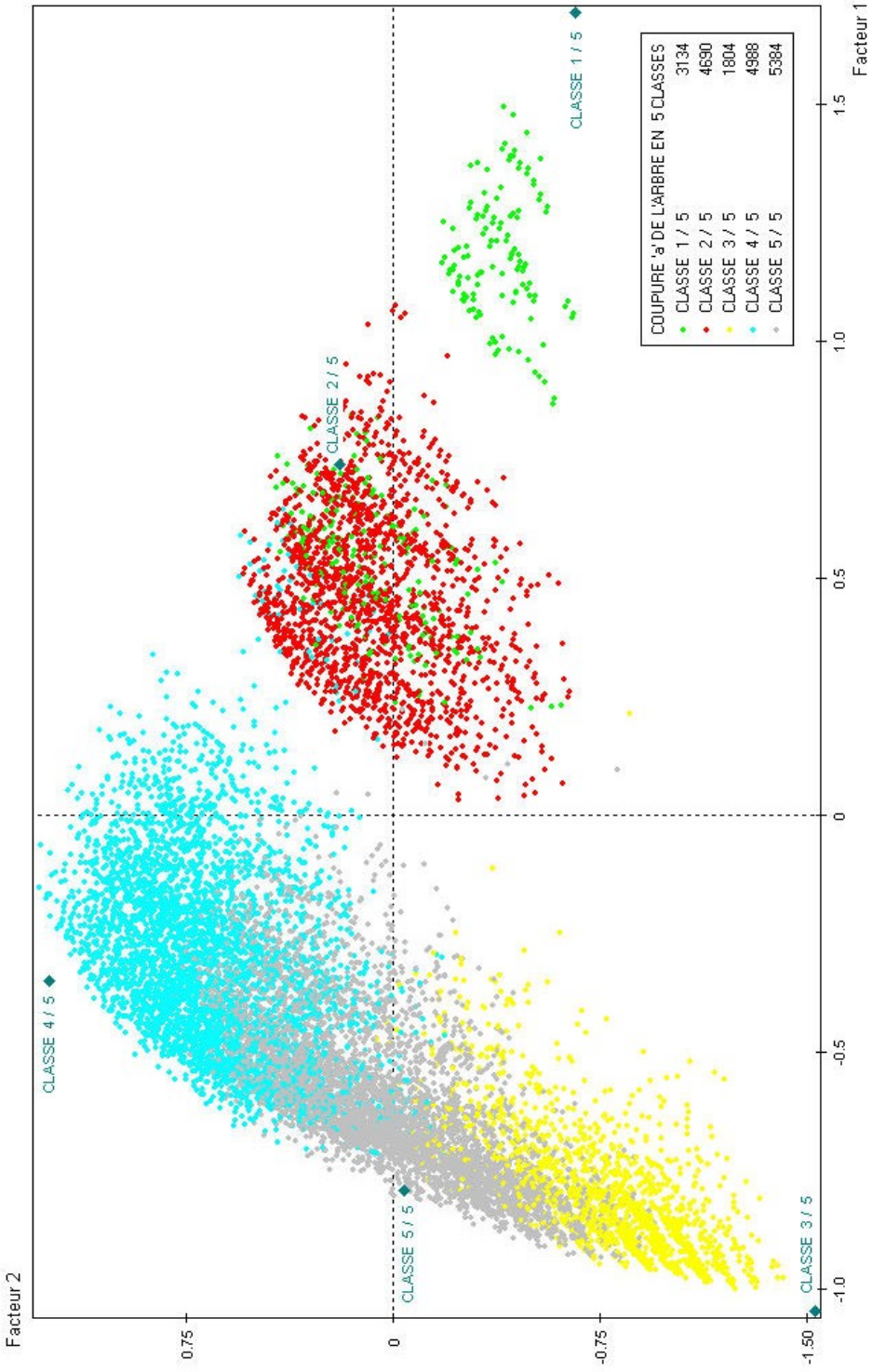


Figure 5.26. Classification sur les indicateurs topologiques, SN00-02 – axes 1 et 2

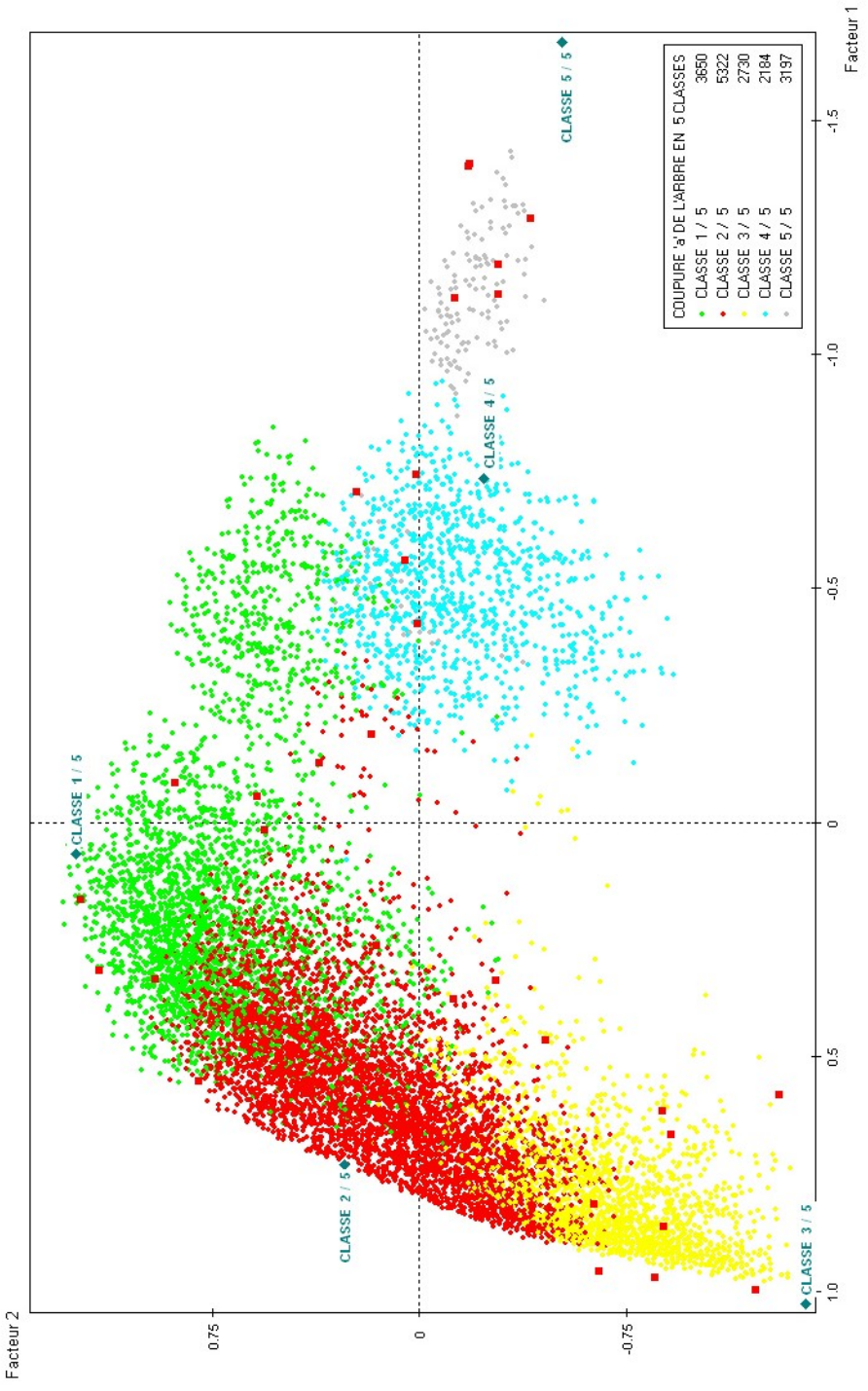


Figure 5.27. Classification sur les indicateurs topologiques, BibUsages – axes 1 et 2

Pour le premier groupe des sessions courtes et rapides, les panélistes de BibUsages paraissent faire preuve de plus d'efficacité que ceux des panels SensNet, avec une part plus importante de parcours éclairs, tandis que les parcours ciblés sont au contraire très peu représentés chez eux (12,8 % des sessions). Il semble que ces utilisateurs avertis vont plus rapidement à l'essentiel lorsqu'ils savent précisément ce qu'ils cherchent. Ceci est corroboré par les entretiens : la plupart des interviewés ont déclaré utiliser les favoris pour classer les sites jugés intéressants, la taille des favoris pouvant varier d'une cinquantaine d'adresses à plus d'un millier. Du côté du panel SN00-02, c'est un mouvement inverse qui se produit, les parcours ciblés étant plus fréquents que les parcours éclairs : la différence avec le panel BibUsages semble indiquer que ce n'est pas l'ancienneté de la pratique qui explique l'efficacité, mais plutôt l'intensité : dans le panel 2000-2002, nous avons pu voir qu'une part non négligeable des internautes sont de faibles utilisateurs, tandis que tous les panélistes de BibUsages sont des usagers réguliers et intensifs du Web.

L'effet d'ancienneté de la pratique semble plutôt paraître au niveau des sessions plus longues : au sein des trois groupes relevant de parcours complexes, c'est dans les parcours à pivot qu'on note les différences les plus notables. Ce type de navigation touche 26,2 % des sessions du panel généraliste 2002, contre 29,1 % des sessions de SN00-02, et 31,1 % de celles de BibUsages. Ce mode de navigation, fortement lié à l'activité de recherche sur le Web à l'aide de moteurs ou de pages-ressources semble attester une maîtrise des outils de recherche : l'utilisateur feuillette les différentes pages dans une logique d'épuisement et de tri de l'offre pour trouver la plus pertinente dans le contexte de sa recherche.

Au cours des entretiens, on a pu voir que cette navigation efficace est opposée par les interviewés à une logique de parcours plus exploratoire et éparse, et qu'elle lui est souvent privilégiée :

> *Je suis un picoreur mais un picoreur qui sait ce qu'il veut. C'est-à-dire que j'essaie de ne pas me, comment dire, de ne pas trop me disperser. Parce que moi, je suis pris par le côté utilitaire ; je suis pas un vagabond, j'aimerais bien mais je n'ai pas le temps. Je me sers d'Internet, en fait comme d'un outil. C'est un outil, c'est un outil, bon ça peut être des fois un outil culturel, donc c'est pour s'amuser, et c'est surtout un outil pour faire des choses, pour trouver de la documentation. (Utilisateur F)*

> *De temps en temps, c'est vrai que je papillonne sur le Web, et c'est vrai que je le fais de moins en moins souvent parce que j'ai moins de temps. […] Je peux être par exemple soit sur le site, par exemple Figaro ou Le Monde ou Libé et en fonction, par exemple, sur le site de Libé, il y a les, les fameux portraits qui sont en dernière page du quotidien et c'est vrai que de temps en temps y a un nom de personnage du portrait qui peut m'intéresser, je peux aller voir son portrait. Ça me renvoie sur une idée et c'est vrai qu'alors soit je note le concept, ou un thème associé et je peux renvoyer sur Google. C'est vraiment de l'hypertexte. (Utilisateur K)*

Le butinage n'est pas rejeté en tant que tel, mais il nécessite un investissement plus important dans la durée, que la plupart ne souhaitent pas assumer : « Je suis pas surfeur, si vous voulez, j'ai pas le temps. » (Utilisateur J).

Pour confirmer ces hypothèses, il est nécessaire de se placer au niveau de l'utilisateur : nous avons envisagé pour l'instant les sessions de manière globale pour chaque panel, de sorte que les utilisateurs intensifs, qui font plus de sessions, pèsent plus lourd dans la représentation finale. En définitive, ces modes de navigation sont-ils liés à l'intensité de la pratique et à l'expertise qui en découle, ou sont-ils partagés par l'ensemble des utilisateurs et déterminés par le contexte local de la tâche ? C'est en examinant la place de chaque type de session pour chaque utilisateur que l'on peut replacer la navigation dans le contexte d'actions normées et examiner la place des déterminations locales et globales dans la morphologie des parcours sur le Web.

> ***Synthèse****. La classification des sessions SN2002 sur la base des indicateurs topologiques fait ressortir cinq parcours-types bien différenciés. D'un côté, parcours éclairs et parcours ciblés forment un groupe homogène de sessions courtes, linéaires ou quasi-linéaires, essentiellement tournées vers les portails généralistes et le WebMail. De l'autre, parcours à détours, parcours à pivots et parcours éclatés suivent une gradation dans la complexité de la navigation, et renvoient à trois contextes d'usage différenciés. Les premiers sont liés aux contenus orientés « vie pratique » et vie hors du Web, et leur linéarité essentiellement rompue par des mouvements courts de* back *; les seconds sont plus apparentés à l'usage de moteurs pour des recherches ouvertes, où certaines pages servent de pivot à la navigation et l'exploration de la Toile ; les derniers, les plus longs et les plus complexes structurellement, sont liés notamment à certains contenus orientés vers les jeux et la communication (WebChat notamment). La projection des sessions des panélistes BibUsages sur ce panorama représentatif des sessions montre la spécificité de ces internautes, plus orientés vers les parcours de recherche et les parcours ciblés.*

Conclusion

Le travail mené sur des données de trafic montre en premier lieu la validité des outils et méthodes de description du contenu et de la forme des parcours que nous avons élaborés. Les premières analyses statistiques sur des données volumineuses et représentatives des usages résidentiels d'Internet en 2002 montrent la grande diversité des comportements de navigation : variété dans les durées, le nombre de sites visités, la complexité des formes de parcours, les thèmes et services accédés, etc. Il en résulte un objet statistique dont la complexité est le reflet de l'inscription des parcours dans des contextes et des pratiques très diversifiés ; ceci justifie une approche par le « geste » plutôt que par le contenu, c'est-à-dire une segmentation des parcours sur la base de leur topologie en première approche. Pour mener à bien cette segmentation, il a été nécessaire de travailler au plus près des indicateurs statistiques et d'opérer des discrétisations *ad hoc*, dans la mesure où les indicateurs que nous

avons construits prennent des significations particulières et entretiennent des corrélations spécifiques pour certaines valeurs.

Ce travail statistique fin nous amène à construire une typologie des parcours Web en cinq classes sur la base de leur forme et de leur temporalité qui rend compte de la diversité des modes de navigation, et des contextes d'usage : les sessions courtes et rapides s'apparentent à des pratiques ciblées où l'utilisateur sait où il va, et s'opposent à des parcours plus diversifiés et plus complexes. Ces éléments morphologiques ne déterminent pas complètement les contenus, mais y sont fortement liés : un comportement exploratoire le restera quel que soit le thème de recherche, mais implique le recours aux outils de recherche et à certains types de pages-ressources qui influencent la forme générale du parcours. Dans les navigations routinières, au contraire, on retrouve des services comme le WebMail ou les informations, qui impliquent une activité répétée et régulière au sein d'un hypertexte connu et maîtrisé. La mise en relation de ces comportements avec le contexte de l'utilisateur en termes d'intensité de pratique et de « territoires personnels » sur le Web doit permettre d'aller plus loin et d'expliquer ces comportements en regard des pratiques individuelles.