

Parcours sur Internet : exploration des traces d'usage

De TypWeb à SensNet

17 septembre 2003

V. Beaudouin – France Télécom R&D

Objectif global : Développer une expertise dans l'analyse des usages d'Internet en croissant :

- Recueil et analyse du trafic Internet
- Qualification des pages vues
- Enquêtes auprès des internautes
- Entretiens/observations in situ

- // études sur la sociabilité téléphonique menées dans le laboratoire UCE, FT R&D

▶ Décomposer l'activité de l'utilisateur derrière son écran pour recomposer des profils d'utilisateurs

Quelques difficultés :

Recueillir le trafic : sondes sur le réseau, sur le poste de l'utilisateur, cookies...?

Qualifier les contenus vus : comment faire à partir d'une simple URL ?

- identifier les types de services utilisés (communication, consultation, recherche, achat...(services utilisés, types de sites consultés)
- Caractériser les pages visitées en terme de forme et de contenu : de quoi parlent les pages visitées, comment en parlent-elles ?

Tenir compte de l'ensemble des protocoles : Web, mail, messageries instantanées, FTP, NNTP...

Objectifs et principes

- ▶ **Objectif : Dresser une cartographie des usages de l'internet**
- ▶ **Principe 0 : L'utilisateur au centre (pas le site)**
 - Seul existe l'internet visité, utilisé (<>Web sémantique)
- ▶ **Principe 1 : Tenir ensemble la production et la réception**
 - la page Web est à la fois un élément d'un site (point de vue de la production) et une composante d'un parcours sur le web (point de vue de la réception)
 - L'url visitée (point de vue usage) dit peu de choses sur le contenu de la page => remonter au contenu et à la structure
- ▶ **Principe 2 : Mettre en relation**
 - la description des sites (observation manuelle et description par des jeux de traits),
 - les parcours des utilisateurs sur les sites
 - le discours de concepteurs et utilisateurs sur leurs pratiques.

De TypWeb (2000-2001)...



Lot 1 : Analyse de la structure et des contenus des sites perso et marchands

- Partenariat FT R&D (V. Beaudouin, M. Pasquier, C. Licoppe), LIMSI (B. Habert), Paris III (S. Fleury)
- Axes de recherche: Mise au point d'une méthodologie d'analyse des pages Web ; comparaison sites marchands / sites perso ; Analyse des sites perso et des pages visitées



Lot 2 : Trajectoires des concepteurs de sites (sites perso musique et marchands) (V. Beaudouin, C. Licoppe)



Lot 3 : Parcours sur Internet

- Partenariat FT R&D, Wanadoo, NetValue, HEC
 - FT R&D : H. Assadi, V. Beaudouin, T. Beauvisage, B. Lelong, C. Licoppe, C. Ziemlicki
 - NetValue : L. Arbues, L. Nicolas
 - HEC : J. Lendrevie

- Axes de recherche : publicité, portails et moteurs, sites marchands, sites personnels, mail, segmentation des internautes

- Méthodes et analyses : développement FT R&D ; Données : NetValue



Lot 4 : Contribution au développement d'une sonde de recueil de trafic (Audinet FTR&D)

=> possibilité de mise en place d'études ad hoc autour de thématiques particulières:

- BibUsages : projet RNRT sur les usages des bibliothèques numériques (H Assadi, T. Beauvisage)
- E-infovalue : utilisation de la messagerie électronique au sein d'un centre de recherche (H Assadi)

À SensNet (2002-2004)



2002-2003 : Projet RNRT SensNet : Catégorisation sémantique des usages et des parcours sur le Web

- Partenaires : France Télécom R&D, NetValue, LIMSI, Paris III
 - FT R&D : H. Assadi, V. Beaudouin, T. Beauvisage, C. Licoppe, C. Ziemlicki
 - Keyrus : A. Bergé
 - NetValue : L. Arbues, L. Nicolas
 - LIMSI : B. Habert, M. Hurault-Plantet, M. Jardino, P. Paroubek
 - Paris III : S. Fleury
- Projet labellisé en mai 2001 ; démarrage 2002



« Données » :

- Traces d'utilisation d'internet récupérées sur le poste utilisateur (≠ approches centrées sur les sites)
- Des cohortes de quelques milliers de personnes représentatives de la population connectée à Internet à domicile
- Descriptif socio-démographique et équipement de chaque internaute
- Une cohorte suivie pendant une année complète (2000, 2001, 2002)
- Prise en compte de tous les usages d'Internet : Web, mail, forums, chat, téléchargement... (≠ approches Web)



Traitements : développement d'une plateforme de traitement et d'analyse des traces d'usage

- Catégorisation des services (différencier services et contenus)
- Exploration des contenus et de la structure des sites (aspiration, normalisation, extraction de traits)
- Exploitation des annuaires du Web (ressources pour catégoriser les contenus)

1. SensNet : Prototype de plateforme de catégorisation

Corpus de sites et de parcours :

- Sélection de sites et de parcours (séquences de pages vues horodatées)
- Aspiration d'un échantillon de sites et de parcours
- Annotation par jeux de traits



Sélection des traits pertinents Traitements statistiques associés



Entretiens auprès
des concepteurs de
sites



Page	Macro- traits	Types d'activité			Traits multimédia				Traits morpho- syntactiques					Traits sémantiques				
		consultation	communication	achat	image	son	video	liens internes et externes	pronoms personnels	structure des phrases	ont	temps employés	catégories des annuaires	catégorisation	ind	catégorie	ajout/suppr	meta-tags
	site 1 institutionnel																	
	site d'association																	
	site de PME																	
n	site marchand																	

Pages Web
(URL)



Processus de validation de la catégorisation, en la confrontant :

- à celle des professionnels
- à celle des internautes

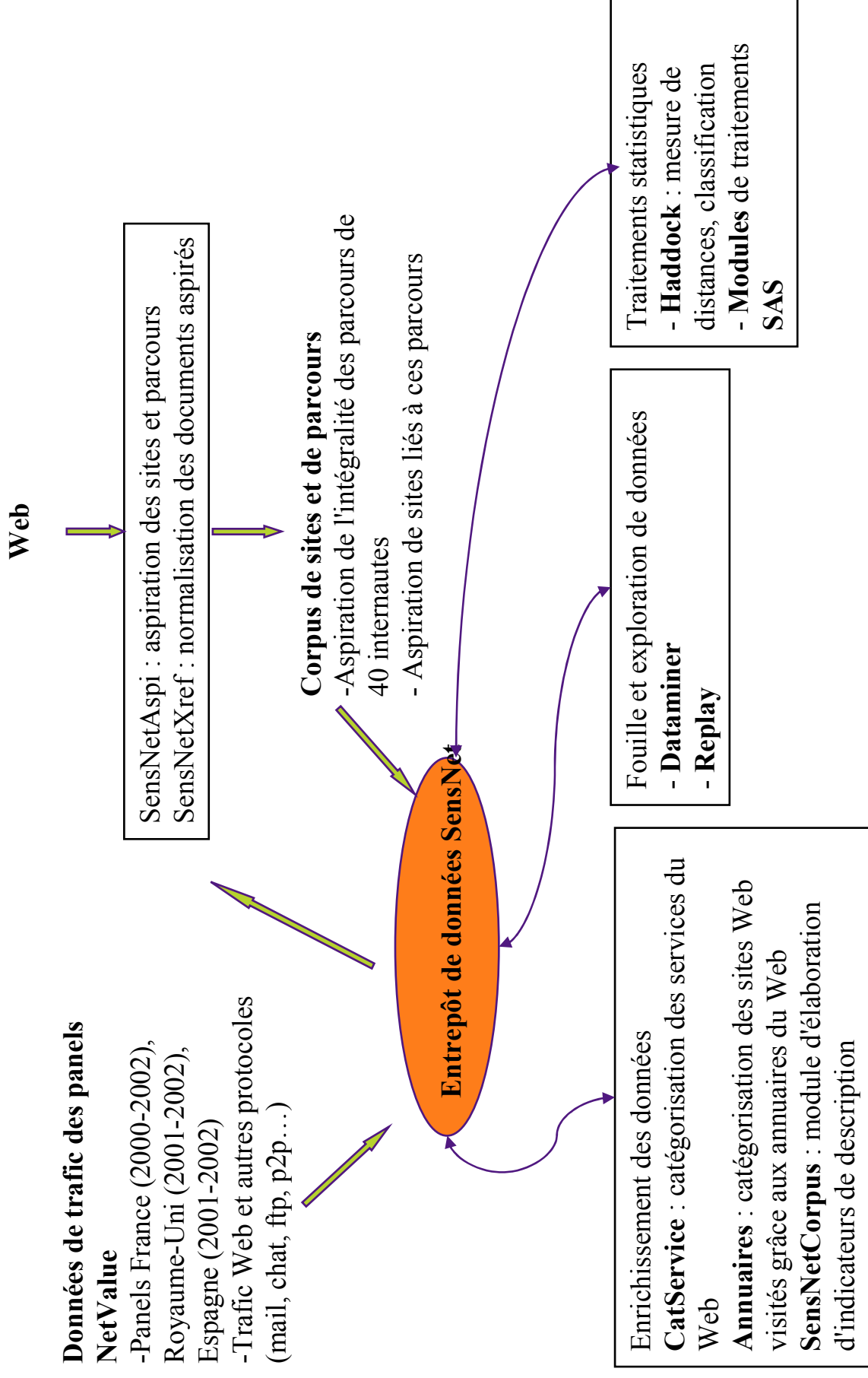


Croisement avec les données de trafic des panélistes (un millier de panélistes x trois pays européens) :

- Catégorisation des sites, projection des parcours des internautes sur les sites.
- Segmentation de parcours et profils utilisateurs ;
- analyse de pratiques spécifiques.



2. SensNet : Représentation fonctionnelle de la plateforme



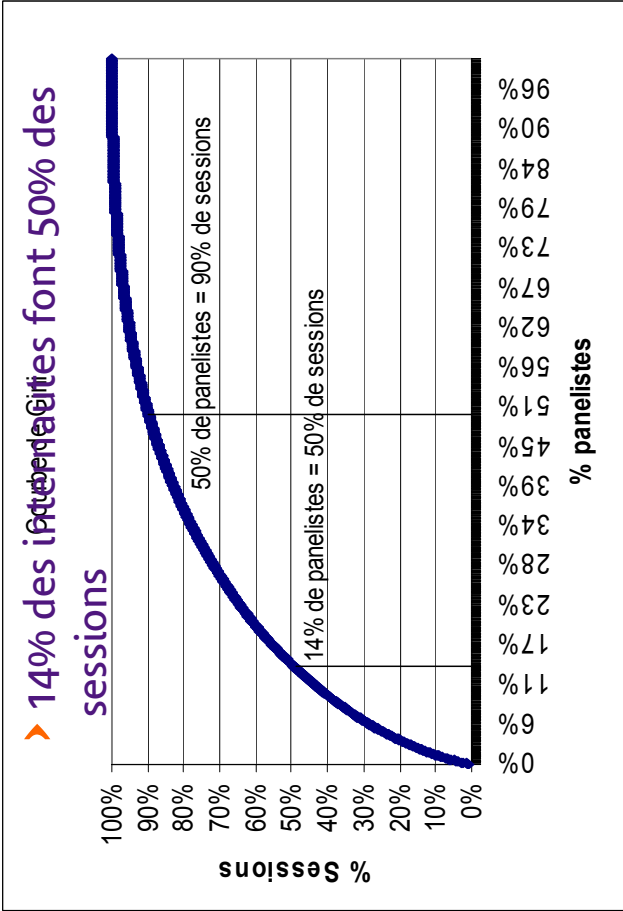
Quelques illustrations



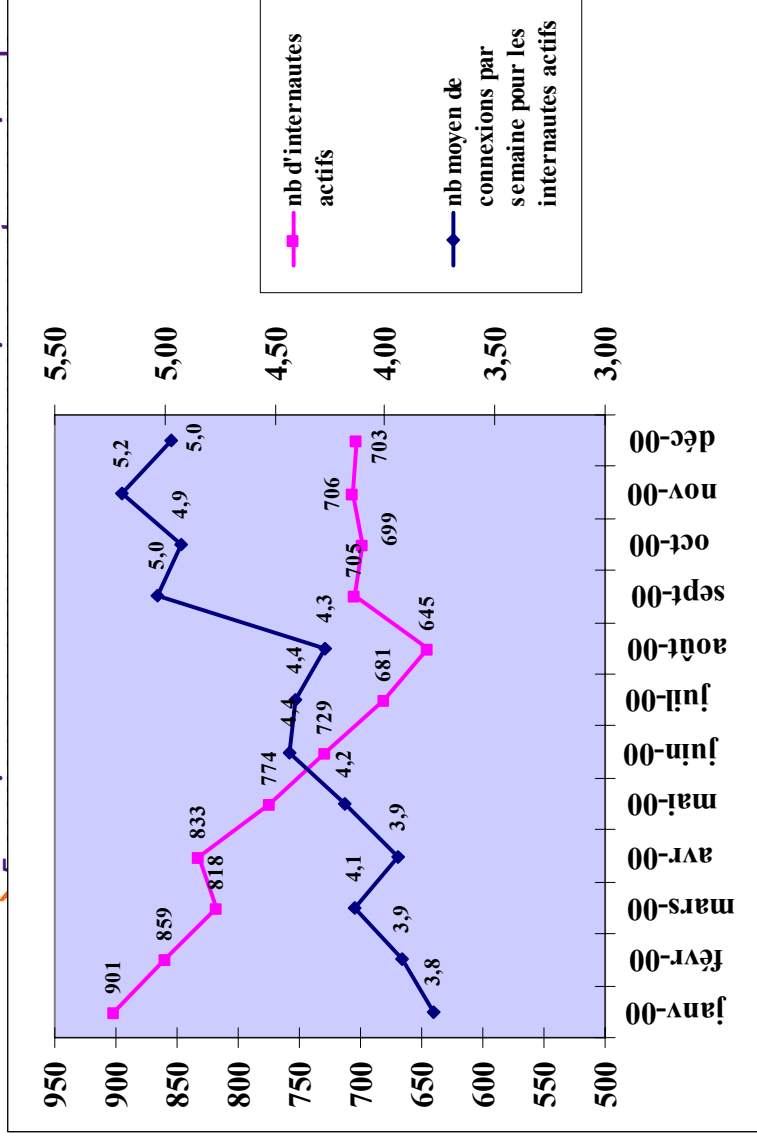
- ▶ Données de cadrage sur les usages d'internet en 2000 :
montrer l'intérêt du suivi longitudinal
- ▶ Segmentation des internautes sur la base de leurs pratiques :
intérêt de l'analyse globale d'internet (web, mail, chat...)
- ▶ Utilisation des moteurs de recherche : catégorisation, fouille
de données et statistique textuelle

Données de cadrage en 2000

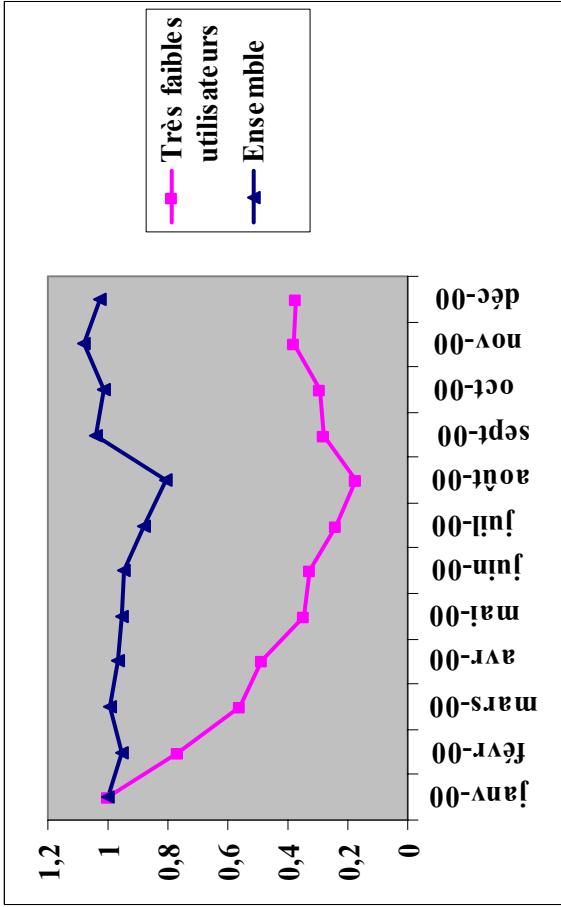
Une distribution très inégale des usages d'Internet



- ▶ Le nombre de sessions Internet est stable au cours de l'année
- ▶ Le nombre d'internautes actifs diminue de mois en mois
 - De 900 à 700
- ▶ Le nombre moyen de sessions pour les internautes actifs progresse



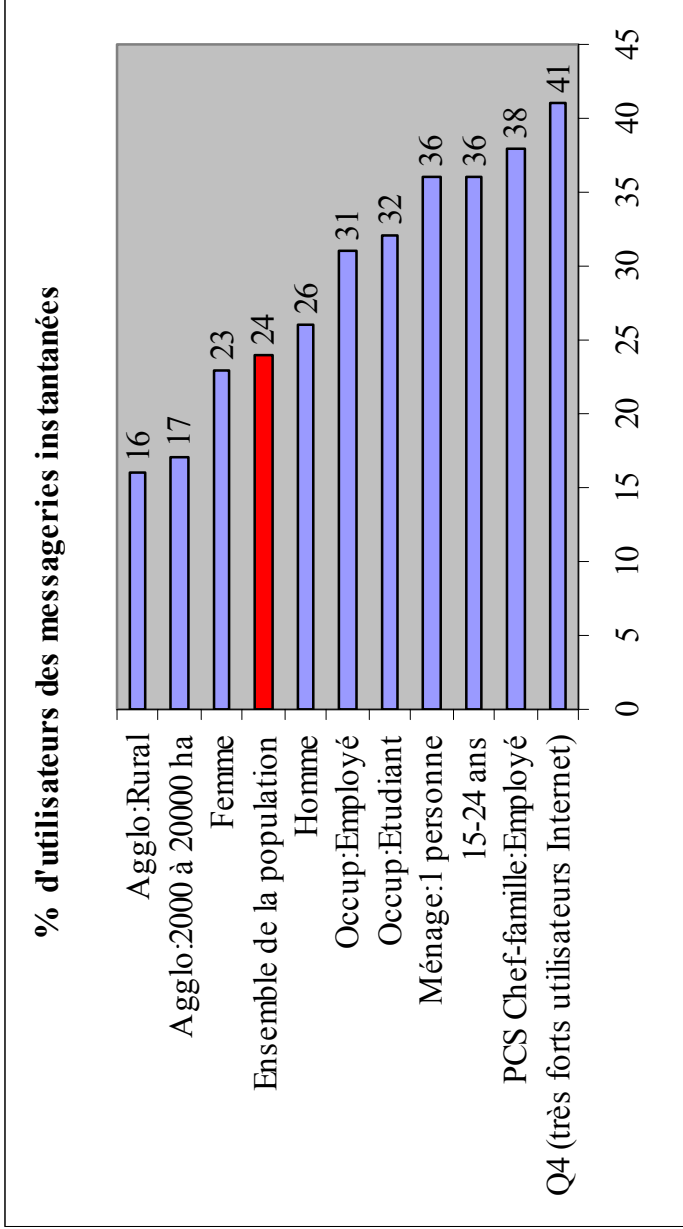
- ▶ Deux trajectoires opposées :
 - les très faibles utilisateurs voient leurs usages décroître
 - Les forts utilisateurs progressent



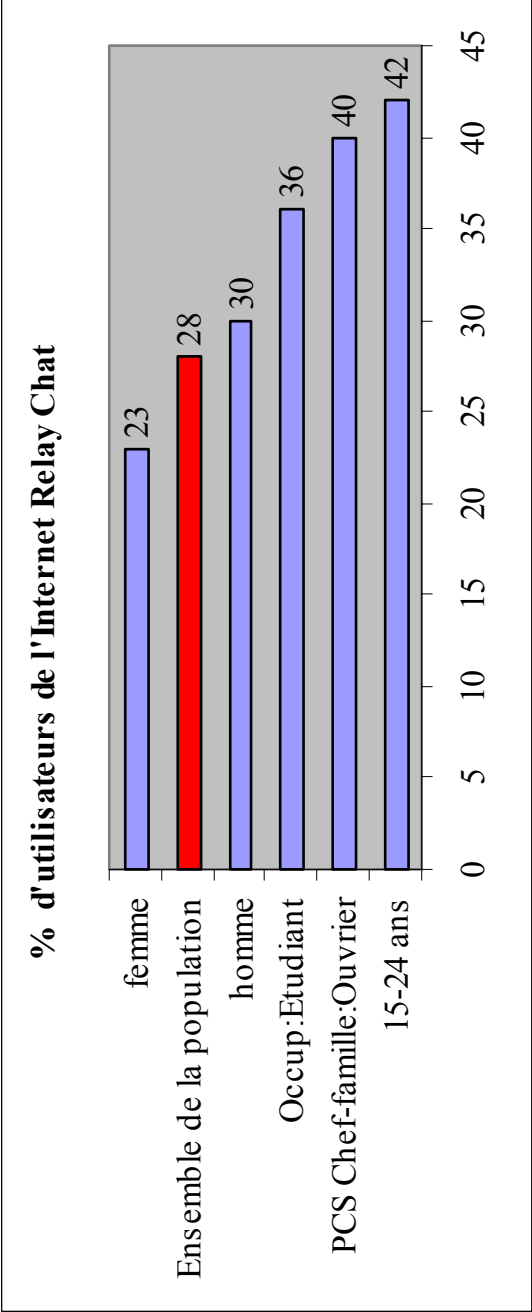
Utilisation des services dans les sessions Internet

Service	% de sessions avec accès au service	Durée moyenne des sessions avec accès au service	% panélistes ayant utilisé au moins une fois le service dans l'année
Mail	38%	23 '	79%
WebMail	13%	37'	74%
Messageries instantanées	11%	36'	25%
Forum (Newsgroups)	5%	31'	24%
Chat	3%	58'	28%
Moteur	17%	43'	93%
WEB	76%	30 '	100%

Messageries Instantanées et Chat



- Facteurs favorisant l'intensité d'usage de l'IM :**
- région parisienne
 - Foyers d'une personne
 - 25-34 ans
 - Occup : ouvrier



- Facteurs favorisant l'intensité d'usage Chat :**
- région parisienne
 - Foyers d'une personne

Construire une typologie des internautes

- ▶ Définir un ensemble de services, protocoles observés
- ▶ Profil des sessions (présence/absence d'un service)
- ▶ Profil des internautes (part des sessions avec accès au service)
- ▶ Analyse en composantes principales et classification ascendante hiérarchique

		Web	Moteur	Mail	WebMail	Chat	ICQ
utilisateur 1	session 1	1	1	0	0	0	0
utilisateur 1	session 2	1	1	0	0	0	0
utilisateur 1	session 3	1	1	0	0	0	0
...							
utilisateur 1140		1	0	0	1	0	1
utilisateur 1140		0	0	1	1	0	1
utilisateur 1140		1	1	0	1	0	1
utilisateur 1140		0	0	1	1	0	1
utilisateur 1140	session 170 000	1	0	0	1	0	1

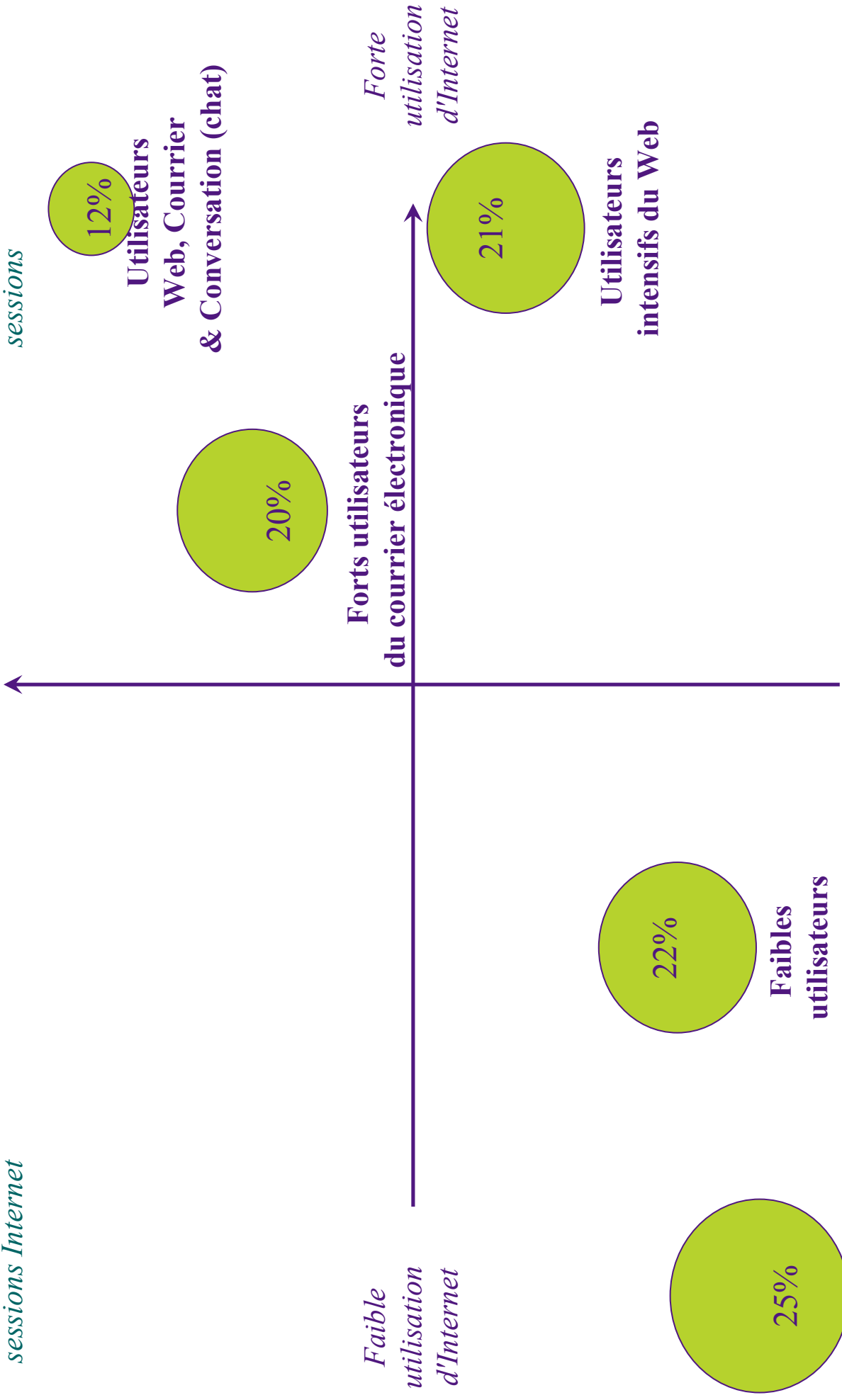
	Web	Moteur	Mail	WebMail	Chat	ICQ	Intensité pratique
utilisateur 1	1	1	0	0	0	0	2
...							
utilisateur 1140	0,6	0,2	0,4	1	0	1	3

Cinq groupes d'internautes

*Faibles utilisateurs : 47%
des internautes et 15% des
sessions Internet*

*Sessions complexes
(Web +mail+chat...)*

*Forts utilisateurs : 53% des
internautes ; 85% des
sessions*



**Très faibles
utilisateurs**

**Faibles
utilisateurs**

*Faible
utilisation
d'Internet*

**Forts utilisateurs
du courrier électronique**

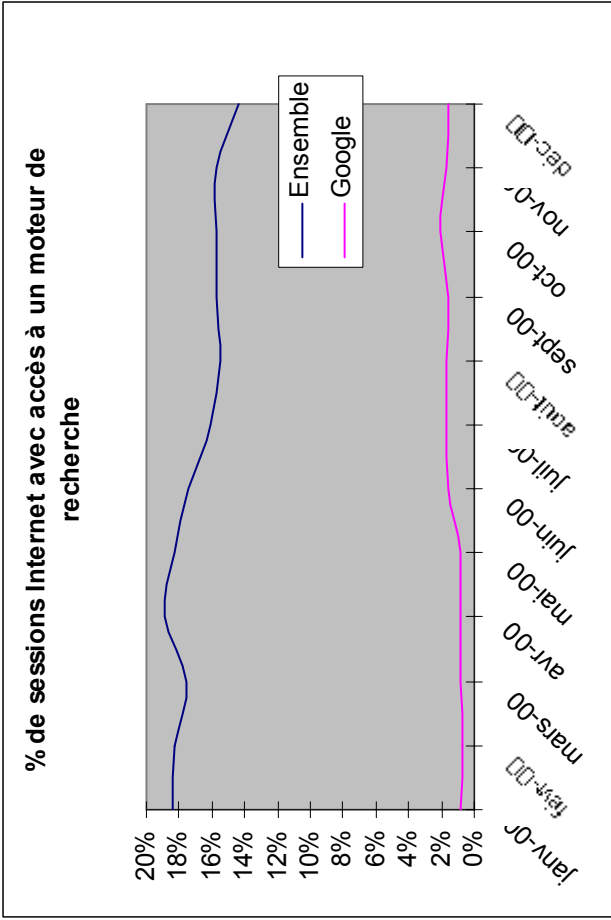
*Forte
utilisation
d'Internet*

*Sessions simples
(Web seul)*

Moteurs de recherche

D'un Internet exploratoire à un Internet balisé

- ▶ La part des sessions Internet sans moteur passe de 82% à 86%
- ▶ Google est le seul moteur en croissance



Typologie des internautes sur la base de leurs requêtes

- ▶ Groupe « culture Internet », recherche d'objets consommables en ligne (images, sons, téléchargement), à la marge de la légalité. Sur-représentation des hommes, des jeunes.
- ▶ Groupe « vie pratique » : recherche d'informations pour une consommation hors ligne. Sur-représentation des femmes, âge moyen.

Requêtes en anglais, jargon technique

Recherche de sites	Recherche thématique
Altavista Download, crack, warez sex, starwars, dongeon Mp3, music	Yahoo
Google	Voilà Sexe, Musique, jeux
Wanadoo Vie pratique : Emploi, people, région, sport, généalogie,	Requêtes en français

Identité des moteurs

- ▶ Confusion entre barre d'adresse et fenêtre du moteur
- ▶ A chaque moteur son public
 - ▶ faible maîtrise des moteurs de recherche
 - ▶ vers un accompagnement des utilisateurs ?
- ▶ Le même moteur Voilà n'est pas utilisé par les mêmes internautes ni de la même manière sur le portail Wanadoo ou Voilà
- ▶ La langue est un facteur de discrimination fort des thèmes de recherche : sexe, piratage, musique... mobilisent davantage l'anglais que la vie pratique.

Quelques avancées, mais...



- ▶ **Un verrou persistant : la catégorisation des pages et des sites vus**
 - Évolution rapide de la technologie des pages/sites, plus rapide que la mise au point des outils d'analyse
 - Aucune "solution" satisfaisante en soi (catégorisation par jeu de règles, en aspirant et analysant les contenus, en exploitant les annuaires) => vers une combinaison de méthodes
- ▶ **Une difficulté à suivre l'internaute dans tous ses usages (les connexions hors du lieu de domicile échappent)**
 - Expériences rares de recueil de traces dans l'entreprise
- ▶ **Un resserrement du marché de la mesure d'audience (trois en un en moins de deux ans)**
 - Plus d'opacité dans les mesures, moins de nécessité de déjouer les "tours" pour gonfler l'audience

Bien du travail en perspective pour l'as...