Usages des moteurs de recherche : une approche centrée utilisateurs

Houssem Assadi, Valérie Beaudouin

France Télécom R&D – DIH/UCE – 38-40, rue du Général Leclerc – 92794 Issy-les-Moulineaux Cedex 9 – France – fax : +33 (0)1 45 29 01 06 – e-mail : {houssem.assadi,valerie.beaudouin}@rd.francetelecom.com

Abstract

Search engines hold an important place in Internet users activity. This is confirmed by both user surveys and web rating services. In this paper, we present the results of a study of search engines uses by a 1140 French Internet users panel, during the whole 2000 year. Our study shows that the use profiles of search engines are distinctly different: the contents indexed by the search engines and those sought by the users vary from a search engine to another. Apparently, Internet users take advantage of this diversity by jointly using several search engines in an exploratory approach or by selecting the search engines that best fit their centers of interest. We also obtained a precise description of the complexity factors of users information seeking sessions and requests.

Résumé

Les moteurs de recherche tiennent une place importante dans l'activité des internautes. Ceci est confirmé aussi bien par les enquêtes auprès des internautes eux-même que par les mesures d'audience publiées par les instituts spécialisés. Nous présentons ici les résultats d'une étude des usages des moteurs de recherche par une population de 1140 internautes français issus du panel NetValue sur l'ensemble de l'année 2000. Cette étude a montré que les profils d'usage des moteurs sont nettement différenciées : les contenus accessibles et recherchées varient d'un moteur à l'autre. Les internautes profitent apparemment de cette diversité en utilisant conjointement plusieurs moteurs dans une approche exploratoire ou en sélectionnant, par « affinité thématique », les moteurs qui correspondent au mieux à leurs centres d'intérêt. Nous avons également obtenu des indicateurs précis sur les facteurs de complexité des sessions de recherche d'information et des requêtes construites par les utilisateurs et avons croisés ces facteurs avec les caractéristiques de la population étudiée.

Mots-clés : Usages d'Internet, moteurs de recherche, analyse de données textuelles, segmentation.

1. Problématique

Les portails font partie des sites les plus visités sur Internet[1]. Ils proposent une gamme plus ou moins étendue de services (recherche d'information, communication, achat…) et jouent un rôle essentiel dans les usages d'Internet en guidant les internautes dans les espaces qu'ils balisent, tout particulièrement au travers de leurs services de recherche de contenus et de services (moteurs de recherche, annuaires et guides généralistes ou thématiques). Ces portails ont en général été élaborés à partir de leur moteur de recherche, et ce service reste la pièce centrale du dispositif[2]. Il faut bien distinguer recherche d'information et utilisation des moteurs : les moteurs de recherche ne sont qu'un des moyens d'accès aux sites sur Internet, au

[1] D'après une étude de Waxman, citée par (DiMaggio *et al.* 2001), ils constituent un quart des sites les plus visités sur le Web.

[2] Nous avons par exemple observé sur le portail voila.fr qu'il y a 7 fois plus de sessions avec un accès au moteur de recherche que de sessions avec la consultation de l'annuaire.

même titre que les annuaires, les liens externes présents dans les sites, les adresses insérées dans les messages (courrier électronique, listes de discussion, forums, chat…), les signets que se constituent les utilisateurs…

Nous nous proposons ici d'explorer les usages des moteurs de recherche sur Internet : les moteurs ont-ils une identité propre ? Comment se distribuent les thèmes de recherche dans la population ? Y a-t-il des associations privilégiées entre des types de recherche et des moteurs particuliers ?

Les résultats présentés ici sont issus d'une étude menée dans le cadre d'un partenariat entre France Télécom R&D, NetValue, HEC et Wanadoo qui portait sur une exploration approfondie des usages d'Internet. Un des axes de recherche de ce partenariat concernait l'utilisation des moteurs de recherche. Nous avons exploité les données de trafic sur Internet d'une cohorte de 1140 internautes du panel français de NetValue sur toute l'année 2000 (soit plus de 7,5 millions de pages vues).

Notre démarche est centrée utilisateur et couvre exhaustivement les moteurs utilisés par notre panel (une trentaine). Dans notre approche des usages, nous associons de manière étroite la statistique descriptive et la statistique textuelle, et nous articulons les données sur l'utilisation des moteurs à l'analyse du contenu des requêtes pour produire une représentation quantifiée et qualifiée des usages. De ce point de vue, notre étude se distingue nettement des études précédentes sur le même sujet, comme le montre l'état de l'art présenté dans la section suivante.

2. Etude des usages des moteurs de recherche : état de l'art

Nous présentons dans cette section une synthèse des différentes études et travaux sur l'utilisation des moteurs de recherche et montrons les spécificités de notre approche.

Les instituts de mesure d'audience sur Internet sont théoriquement capables de fournir des résultats sur l'utilisation des moteurs de recherche. Mais si l'on examine les études d'audience publiées, il en ressort que la fonctionnalité « moteur » n'est pas précisément identifiée et comptabilisée sur les grands portails offrant un service de recherche d'information sur Internet (tels que Yahoo!, MSN, AOL, Lycos, etc.). Par exemple, sur un portail comme yahoo.com, qui offre une palette assez large de services (moteur de recherche, annuaire du Web, actualités, WebMail, Chat…), l'audience calculée ne permet pas de faire un décompte précis de l'usage de tel ou tel service (dont le moteur de recherche). Il en résulte que les palmarès des moteurs de recherche publiés par ces instituts[3] ne donnent pas une image correcte de la réalité des usages des moteurs par les internautes.

Les enquêtes quantitatives auprès des internautes soulignent l'importance de la recherche d'information, qui figure systématiquement dans le groupe de tête des activités citées[4], mais elles ne permettent pas d'étudier spécifiquement l'utilisation des moteurs.

Par ailleurs, plusieurs équipes de recherche universitaires ou industrielles se sont intéressées à la question des usages des moteurs, en exploitant le contenu des requêtes. La plupart de ces études sont « centrées moteur », c'est à dire qu'elles s'intéressent à un seul moteur, les données exploitées étant les fichiers de traces (*logs*) disponibles sur le serveur (cf. (Jansen &

[3] Voir les rapports des différents instituts à ce sujet, regroupés sur le site de « Search Engine Watch » : http://searchenginewatch.com/reports/index.html

[4] Voir notamment les résultats d'une enquête récente en France dans (Heitzmann & Loué, 2001).

Pooch 2000) pour une revue récente de ce type de travaux). Ces travaux présentent trois limites. Premièrement, le fait de travailler sur un seul moteur de recherche pourrait présenter un biais non négligeable : il n'est pas acquis que les usages d'un seul moteur soient représentatifs des usages de tous les moteurs (cf. nos résultats sur les profils des moteurs § 4). Deuxièmement, ces travaux basés sur les traces recueillies sur les serveurs hébergeant les moteurs ignorent tout des caractéristiques des utilisateurs des moteurs. Enfin, ces études utilisent des données issues d'une période d'observation relativement courte (de 1 à 42 jours). Or nous avons observé une saisonnalité des recherches, immanente ou liée à l'actualité.

Les travaux « centrés utilisateurs » sont quant à eux plus rares. Ce sont des études expérimentales utilisant essentiellement des méthodes de la psychologie cognitive et de l'ergonomie (expériences « en laboratoire », observations, modélisation de l'activité), et travaillant à partir de populations restreintes et non représentatives. Ces travaux ont un intérêt majeur, ils permettent de comprendre les stratégies de recherche d'information, ce qui va au-delà de l'utilisation des moteurs (d'autres moyens sont utilisés par les internautes pour accéder à des sites), et d'obtenir une appréciation qualitative des résultats fournis par le moteur. En combinant mesures et enquêtes auprès des utilisateurs, ces études permettent de modéliser l'activité de recherche d'information (Hölscher & Strube, 2000), de comprendre les effets de l'expertise (dans un domaine et dans la recherche sur Internet) sur les pratiques de recherche (Hölscher & Strube, 2000), de comprendre l'impact des fonctionnalités avancées de recherche d'information sur la pertinence des résultats (Bruza *et al.* 2000), d'apprécier la qualité des pages de résultats des moteurs par rapport à la tâche à effectuer (Amitay & Paris, 2000).

Notre approche est innovante à plus d'un titre. Il s'agit tout d'abord d'une cohorte d'un millier de personnes. La cohorte, issue du panel NetValue, rend compte de la diversité de la population connectée à Internet[5] : les internautes sont qualifiés par leur profil socio-démographique, leur équipement et leurs usages d'Internet. La durée d'observation longue (toute l'année 2000), inédite dans ce domaine d'investigation, nous permet d'étudier l'évolution des usages des moteurs au fur et à mesure de l'apprentissage et de l'appropriation, de qualifier les utilisateurs selon leur intensité d'usage, selon les moteurs utilisés… Cette observation prolongée permet de tenir compte des variations des thèmes avec le temps et de collecter un échantillon de requêtes représentatif des sujets de recherche des utilisateurs sur toute l'année. Enfin, nous tenons compte de l'ensemble des moteurs de recherche utilisés, une trentaine, ce qui nous permet de connaître le positionnement relatif des moteurs. Enfin, nous articulons dans notre analyse des méthodes de statistiques traditionnelles, y compris l'analyse des données, avec les méthodes de statistique textuelle : cette combinaison permet de construire une représentation qualifiée des usages.

3. Les données

Dans le cadre du partenariat décrit précédemment, NetValue a mis à notre disposition des données issues de son panel d'internautes français[6], soit une cohorte de 1140 internautes suivie sur l'ensemble de l'année 2000.

[5] Le panel NetValue est recruté selon des critères stricts de représentativité. L'étude porte sur une cohorte suivie sur un an, période longue pendant laquelle la structure des internautes peut avoir changé. La cohorte, qui était représentative des internautes à domicile le premier mois, devient donc moins représentative à la fin de l'année.

[6] Voir http://www.netvalue.fr/ pour les détails sur les panels d'internautes de NetValue.

3.1. Terminologie

Des termes tels que « session » ou « requête » sont très utilisés dans la littérature traitant des usages d'Internet sans être systématiquement et clairement définis. Nous donnons ici nos propres définitions des termes principaux que nous utilisons.

Session : une session est constituée par la suite des adresses (URL) des pages Web visitées par un utilisateur donné. Les frontières des sessions sont déterminées en fonction d'une durée d'inactivité de 30 mn : si l'utilisateur n'a pas demandé une nouvelle URL pendant une durée supérieure ou égale à 30 mn, nous considérons que la session en cours est terminée.

Requête : une requête vers un moteur de recherche est une requête adressée par un utilisateur à un moteur de recherche donné, pendant une session donnée avec un ou plusieurs mots-clefs donnés. Si un utilisateur adresse à plusieurs reprises une requête composée par les mêmes mots-clefs au même moteur pendant la même session, nous comptabilisons une seule requête. En revanche, si l'utilisateur utilise les mêmes mots-clefs avec le même moteur dans une autre session, nous comptabilisons une nouvelle requête.

Mot clef (syn. terme) : il s'agit d'une suite de caractères ne contenant pas de blancs et apparaissant dans le champ prévu à cet effet dans la requête. Les opérateurs (voir infra) ne sont pas comptabilisés comme mots-clefs.

Opérateur : il s'agit de signes permettant de relier les mots-clefs de la requête pour construire une expression complexe. Nous tenons compte des opérateurs suivants : AND, OR, NOT, NEAR, +, -, *, &, ". Remarque : certains moteurs de recherche francophones permettent de désigner les opérateurs booléens par leur nom français (ET, OU, NON), nous n'avons pas traité ce cas.

3.2. Méthode de reconnaissance des requêtes moteur

Le premier problème qui s'est posé dans notre étude est celui de l'identification, parmi l'ensemble des URL vues par les internautes du panel, de celles correspondant à une requête adressée à un moteur de recherche.

Nous avons mis au point un système à base de règles – représentées sous forme d'expressions régulières – qui permet d'identifier de manière sûre toutes les URL qui correspondent à l'affichage d'une page de résultats d'un moteur de recherche. Le système de règles repose sur le nom du site (par exemple : search.voila.fr, google.yahoo.com) et sur la présence d'un certain nombre de séquences qui permettent d'identifier de manière sûre que l'URL correspond bien à une requête moteur (par exemple : la séquence 'kw=' pour Voila).

Suite à l'identification des pages de résultats, un module se charge du transcodage (transformation des codes du type %E9 en caractères spéciaux, etc.) et de la reconnaissance des mots-clefs composant la requête et des opérateurs utilisés.

Exemple :
URL de départ :
http://www.euroseek.net/query?ifl=uk&query=photoshop+AND+6+AND+t%E9l%E9charger&domain=world&domain=world&domain=world&lang=fr
Requête transcodée : photoshop AND 6 AND télécharger
Mots-clefs : photoshop 6 télécharger
Opérateurs : AND (2)

Ensuite, en respectant notre définition de la requête, nous identifions parmi les pages de résultats celles qui correspondent effectivement à une requête. Les traitements ne portent ensuite que sur les requêtes ainsi définies.

3.3. Statistiques globales concernant l'usage des moteurs

L'usage des moteurs concerne une large partie des internautes, en effet, 93% (1055 sur 1140) des panélistes ont effectué au moins une requête moteur sur l'ensemble de l'année 2000. Ces internautes ont fait appel à 29 moteurs de recherche différents. Plus de 100 000 requêtes ont été émises, ce qui ne représente que 1,3% des 7,5 millions de pages vues par nos panélistes sur l'ensemble de l'année. Mais la couverture en termes de sessions est plus importante, en effet, 20% de sessions Web contiennent au moins une requête moteur (ce qui représente environ 30 000 sessions). Nous avons également établi qu'une session comportant l'utilisation d'un moteur dure en moyenne 30 minutes, alors qu'une session Web ne dure en moyenne que 16 minutes. L'utilisation des moteurs semble rallonger la longueur des sessions.

L'intensité d'usage des moteurs est corrélée avec l'intensité d'usage d'Internet, et reflète donc des usages avancés d'Internet. L'usage des moteurs, reproduit en les aggravant les inégalités d'usage observées sur Internet (les femmes les utilisent nettement moins que les hommes, les personnes âgées que les jeunes…).

4. Les moteurs ont des identités marquées

Nous avons cherché à caractériser le profil de chacun des moteurs, en fonction des caractéristiques des panélistes qui les utilisent et du contenu des requêtes qui leur sont adressées. Nous avons regroupé les requêtes en fonction du moteur utilisé, et cherché à identifier les caractéristiques des internautes auteurs de ces requêtes, caractéristiques socio-démographiques mais aussi caractéristiques en termes d'usages d'Internet : autrement dit, nous identifions les modalités de variables spécifiques d'un moteur donné.

Ensuite pour identifier les types de requêtes spécifiques de chaque moteur, nous avons constitué un corpus dans lequel chaque moteur est décrit par l'ensemble des mots-clefs des requêtes qui lui ont été adressées. Nous avons calculé le vocabulaire spécifique de chaque moteur à l'aide des logiciels Alceste conçu par Max Reinert (1993) et Lexico conçu par André Salem (Lebart & Salem, 1994). Les deux outils se complètent et utilisés conjointement permettent d'affiner l'interprétation[7].

Les moteurs qui possèdent les identités les plus marquées et les plus opposées sont *Altavista* et *Wanadoo.*

Altavista est un moteur essentiellement utilisé par des hommes (88% des requêtes y sont faites par des hommes contre 79% pour la moyenne) et par d'anciens internautes (42% des requêtes sont faites par des personnes connectées avant 1998, contre 32 % en moyenne). En ce qui concerne les requêtes, *Altavista* recueille beaucoup de requêtes en anglais, et les recherches portent sur des objets consommables sur Internet : informatique (logiciels, utilitaires…), musique (fichiers mp3 et paroles de musique), sexe (photos, vidéo…) et jeux. Les recherches portent sur des produits gratuits et il n'est pas étonnant que ce moteur soit également utilisé pour des recherches à la marge de la légalité : recherche de logiciels piratés, téléchargement de morceaux de musique en mp3…

A l'inverse, *Wanadoo* est un moteur plus féminin (28% de requêtes de femmes contre 21% en moyenne), d'internautes plus récents (47% connectés en 1999 contre 39% en moyenne) et

[7] Alceste et Lexico n'utilisent pas les mêmes méthodes pour calculer le vocabulaire spécifique et tandis qu'Alceste s'appuie sur du vocabulaire lemmatisé par des modules propres au logiciel, Lexico s'appuie sur les formes brutes.

plus provinciaux. Les requêtes adressées à *Wanadoo* se caractérisent par la forte présence de recherche de sites : ce sont des noms ou adresses de sites qui sont tapés dans la fenêtre du moteur.

Le moteur *Yahoo* est utilisé par des jeunes (35% des requêtes sont faites par des 15-24 ans contre 25% en moyenne) et plutôt par d'anciens internautes (39% connectés avant 1998 contre 32% en moyenne). *Voila* incarne assez bien le profil moyen des moteurs, et a donc un positionnement peu spécialisé en ce qui concerne le profil socio-démographique de ses utilisateurs. En terme de requêtes, *Yahoo.fr* et *Voila* se situent entre les deux pôles définis par *Altavista* et *Wanadoo* : les moteurs sont assez proches, puisqu'ils recueillent tous deux des requêtes sur la vie pratique ; mais *Voila* est plus spécialisé dans l'emploi, le logement, le commerce, tandis que *Yahoo.fr* recueille des requêtes sur les jeux, les études, le sexe, ce qui est sans doute un effet de génération. De ce point de vue, *Yahoo.fr* est plus proche d'*Altavista* que ne l'est *Voila*. *Yahoo.com* est quant à lui encore plus proche d*'Altavista* : requêtes en anglais portant sur des objets propres à l'univers d'Internet.

Nous avons montré que chaque moteur a des traits de caractère qui le distingue des autres. La personnalité du moteur est révélée par le profil socio-démographique de ceux qui l'utilisent et par le type de requêtes qui lui est adressé. Les moteurs ont chacun des caractéristiques propres (mode d'affichage des résultats, couverture, type de sites indexés, fréquence des mises à jour…), un style, qui rencontrent des types d'usages spécifiques. On peut faire l'hypothèse qu'au fil du temps s'accroît l'adéquation entre le moteur et ses utilisateurs. Si Altavista recueille beaucoup de requêtes sur le téléchargement de musique et de logiciels, c'est sans doute qu'il référence mieux que les autres ce type de site, ce qui conduira des auteurs de sites de cette catégories à privilégier le référencement chez Altavista. En ce sens, il y a bel et bien une co-construction entre l'offre et la demande sur les moteurs.

Examinons à présent quels sont les profils des internautes en fonction des requêtes qu'ils adressent aux moteurs.

5. Typologie des utilisateurs selon les thèmes de recherche

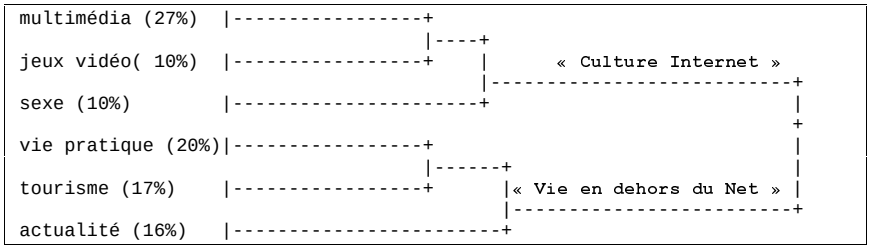
Partant du constat que les requêtes les plus populaires ne représentent qu'une infime partie de l'ensemble des requêtes adressées aux moteurs[8], nous avons choisi d'adopter une approche contrastive afin de capter ce qui différencie et distingue les groupes de requêtes et, par extension, les groupes d'internautes ayant engendré ces requêtes. Pour illustrer la prédominance des requêtes rares, voici quelques chiffres extraits de nos données : la requête la plus fréquente tous moteurs confondus (« sexe ») ne représente que 0,3% de l'ensemble des requêtes. Les requêtes ayant une fréquence inférieure ou égale à 3 représentent 50%, et celles ayant une fréquence de 1 représentent 33% de l'ensemble des requêtes. Etant donné que les requêtes singulières (celles n'apparaissant que peu de fois dans les données) sont les plus intéressantes, seule une approche fondée sur la classification, c'est à dire le regroupement des mots-clefs dans des ensembles homogènes, peut permettre de concrétiser une analyse thématique des requêtes adressées aux moteurs.

L'objectif de cette partie de l'étude est de constituer des groupes d'utilisateurs homogènes par rapport aux thèmes de leurs requêtes sur les moteurs. Le principe est de rapprocher deux

[8] Ce constat a été rapporté dans plusieurs études et confirmé à partir d'importants volumes de données constituées à partir des traces (*logs*) de moteurs de recherche. Voir par exemple (Jansen *et al.* 2000). (Lajoie 2001) présente également des éléments allant dans ce sens.

individus lorsqu'ils utilisent, de manière significative par rapport à l'ensemble de la population, les mêmes mots clefs dans leurs requêtes. Pour cela, nous avons créé un fichier de données où chaque utilisateur était caractérisé par l'ensemble des requêtes qu'il a adressé à un moteur de recherche durant l'année 2000. Nous avons ensuite soumis ce fichier à l'outil d'analyse de données textuelles Alceste (Reinert, 1993). Enfin, pour caractériser les classes proposées par Alceste, nous avons utilisé une procédure statistique qui permet de dégager les variables caractérisant la sous-population d'utilisateurs de chaque classe. Ces variables sont de deux types : (1) variables socio-démographiques, (2) variables d'usage (intensité d'usage du mail, du Web, …).

Alceste propose une classification des panélistes utilisateurs de moteurs selon l'arbre ci-dessous.



Clef de lecture : Alceste construit automatiquement 6 classes, qui sont regroupées à un niveau supérieur en deux grandes classes. La première classe (en haut à gauche) représente 27% du corpus des requêtes, nous l'avons intitulée « multimédia ».

Figure 1 – Classification des panélistes à partir de leurs requêtes moteur

La classe que nous avons intitulée « culture Internet » regroupe des requêtes liées au « monde Internet » : multimédia (photo, vidéo, musique), sexe, jeux vidéo et piratage de logiciels. Les utilisateurs ayant émis ces requêtes sont plutôt masculins, jeunes (âgés de moins 24 ans), étudiants, et n'ayant pas ou peu de revenus. Au niveau des usages d'Internet, cette population se caractérise par une très grosse consommation de pages Web associée à un usage plutôt réduit du mail.

La classe que nous avons intitulée « la vie en dehors du Net » regroupe des requêtes liées à la vie pratique : tourisme, actualités, santé/beauté, ... Il s'agit d'une population plutôt féminine, plus âgée que la précédente (de 35 à 49 ans), avec une présence forte des professions intermédiaires. Au niveau des usages d'Internet, cette population se caractérise par une très grosse consommation de pages Web et un usage très important du mail.

A un deuxième niveau d'analyse, la partition ci-dessus peut être affinée. Nous décrivons ici la subdivision de la classe « culture Internet » en trois sous-classes :

1 La première regroupe des requêtes liés aux thèmes des logiciels, du piratage, du téléchargement et du multimédia, avec des mots-clefs comme : informatique, crack, download, DVD, free, logiciel, windows, linux, mp3, …

2 La deuxième regroupe des requêtes autour du sexe, on y trouve également le thème de la gratuité.

3 La troisième regroupe des requêtes autour des jeux vidéo. Les mots-clefs les plus caractéristiques de cette classe correspondent à des noms de jeux vidéo, à des synonymes du terme « astuce », ainsi qu'à des noms de marques et d'entreprises dans le secteur des jeux vidéo.

6. Typologie des internautes selon leurs profils de session

Précédemment, nous avons identifié des types de moteurs et des types d'internautes en fonction des requêtes. Tout serait infiniment simple si les internautes se contentaient de n'utiliser qu'un seul moteur pour effectuer leurs recherches sur Internet et n'avaient qu'un seul centre d'intérêt. Il semble au contraire que le recours à différents moteurs, au fil du temps, mais aussi au sein même d'une session Internet soit une pratique courante, et que ce soit même un signe d'usages avancés des moteurs. Les trois-quarts des panélistes ont utilisé au moins trois moteurs différents dans l'année (20% en ont utilisé plus de 10 !) et dans 32% des sessions au moins deux moteurs différents sont utilisés.

Il nous a paru donc indispensable de construire une typologie des parcours de recherche, en fonction des moteurs utilisés, avant de bâtir une typologie des internautes fondée sur le profil de leurs sessions.

Pour construire une typologie des sessions de recherche, nous avons tout d'abord isolé les sessions dans lesquelles un seul moteur était utilisé. Pour les sessions faisant intervenir plus d'un moteur, nous avons construit une typologie avec une classification hiérarchique ascendante calculée sur les résultats d'une analyse des correspondances multiples. Ces traitements permettent d'identifier les profils de sessions. Deux types de profils émergent : des profils de sessions monomoteurs et des profils multimoteurs.

Profils de sessions	**effectifs**	**%**
Altavista	2779	10
Wanadoo-Voila	9681	33
Yahoo (.fr et .com)	7225	25
Autres moteurs utilisés seuls	4487	15
Multimoteurs	4904	17

Tableau 1. Profil des sessions

Le Tableau 1 présente un regroupement des sessions en fonction du moteur dominant. Les sessions avec *Voila* et *Wanadoo* représentent un tiers des sessions. Un quart des sessions mobilisent exclusivement *Yahoo.fr* et *Yahoo.com*, 10% des sessions ont *Altavista* comme moteur unique ou principal. Ensuite, nous avons regroupé les sessions où un seul moteur (autre que *Yahoo*, *Voila* ou *Altavista*) est utilisé, ce qui représente 15% des sessions. Enfin un dernier groupe, 17%, rassemble les sessions ou plusieurs moteurs différents sont utilisés. Nous avons pu évaluer que 7,5% des sessions avaient recours à un métamoteur[9]. Il y a donc 10% des sessions où plusieurs moteurs sont utilisés en parallèle sans outil dédié.

Sur la base des profils de sessions détaillées, nous avons construit une typologie des utilisateurs de moteurs qui permet d'identifier six groupes. Les quatre premiers sont marqués par une prédominance forte d'un moteur. Pour le premier groupe, qui regroupe 27% des internautes, il s'agit des moteurs *Wanadoo* et *Voila.* Ils se distinguent des autres par un accès à Internet assez récent, un profil classe moyenne et une sur-représentation des femmes. Le second groupe, 18%, regroupe des utilisateurs de *Yahoo*, surtout des étudiants, parisiens et plutôt anciens internautes. Le troisième groupe, 7% des internautes, utilise *Altavista* de manière privilégiée : ce sont surtout des hommes et d'anciens internautes. Les internautes du quatrième groupe, 16%, utilisent un seul moteur de façon quasi exclusive. Ce sont de très

[9] Pour identifier les requêtes méta-moteur, on a posé qu'il fallait qu'une même requête soit adressée à au moins 3 moteurs et qu'il y ait un délai inférieur à 90 secondes entre la première et la dernière page de résultats fournie.

faibles utilisateurs des moteurs. Le cinquième groupe se distingue par l'importance des sessions multimoteurs : on y trouve à la fois d'intenses utilisateurs des métamoteurs et des internautes récents qui les utilisent rarement. Enfin, le sixième groupe, qui représente un quart des internautes, a un profil de sessions proche de la moyenne : ces derniers ont utilisé tous les moteurs et toutes les combinaisons de moteurs au fil de l'année. Ce sont d'intenses utilisateurs des moteurs, surtout des hommes, des Parisiens, et d'anciens internautes.

Il semble que l'intensification de l'usage des moteurs passe par une diversification des moteurs utilisés. Les faibles utilisateurs n'utilisent quant à eux qu'un seul moteur. Il est probable que ce résultat ne se maintienne pas en 2001. En effet, il semble que l'utilisation de Google s'accompagne d'une fidélisation accrue à ce moteur.

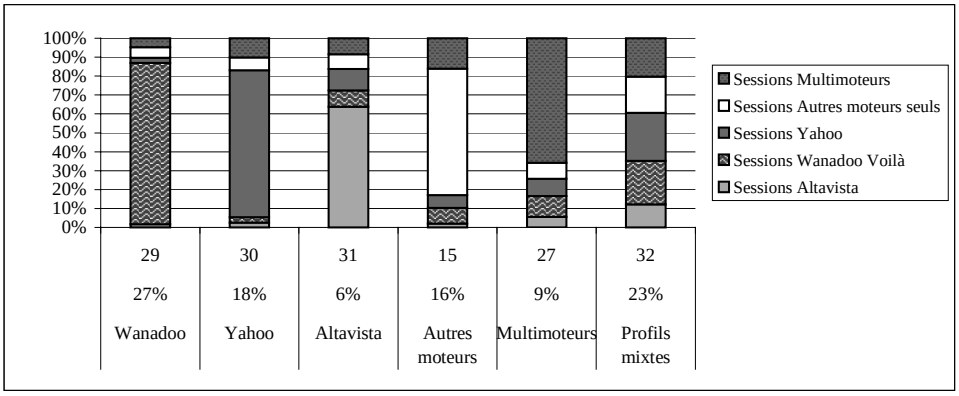


Figure 2. Typologie des panélistes sur la base de leurs profils de sessions

Voyons maintenant, pour les internautes, s'il existe des liens entre les types de moteurs utilisés et le type de requête. Le tableau suivant montre qu'il y a bien des affinités entre les types de moteurs utilisés et les thèmes de recherche privilégiés des internautes. Les internautes qui utilisent principalement *Altavista* sont nettement plus nombreux que la moyenne à appartenir au groupe des internautes classés dans « Culture Internet ». Inversement, les utilisateurs exclusifs de *Wanadoo-Voila* sont plus nombreux dans « Vie pratique », tout comme ceux qui utilisent un seul moteur de manière exclusive.

Ces statistiques montrent que les internautes mettent à profit l'offre pléthorique de moteurs de recherche sur Internet. Pour les internautes utilisant beaucoup de moteurs différents (profil mixte et multi-moteur), il nous reste à montrer qu'ils choisissent leur moteur en fonction du type de requête. C'est une des perspectives de recherche qui émerge de ces premiers résultats.

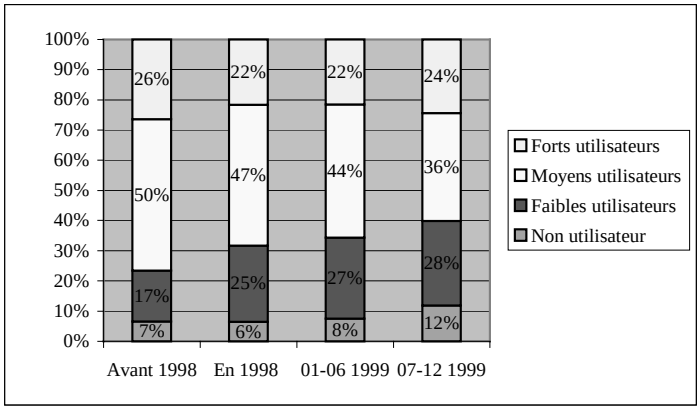
	Wanadoo	Yahoo	Altavista	Autres moteurs	Multimoteur	Mixte	Total
Vie pratique	72	54	35	64	56	56	60
Internet	28	46	65	36	44	44	40

Tableau 2. Répartition des internautes par moteur et par classe thématique de requêtes

7. Evolution de l'usage des moteurs et de la complexité des requêtes

Dans cette partie de l'étude, nous nous sommes intéressés à la complexité des requêtes construites par les internautes et à leur évolution dans le temps. Nous avons lié cette question au processus d'appropriation des moteurs par les utilisateurs et nous avons croisé ces données avec l'ancienneté des internautes. Nous nous situons dans un contexte où globalement on observe une diminution de l'usage des moteurs au fil de l'année.

Nous avons d'abord noté que le lien entre l'ancienneté d'utilisation d'Internet et l'intensité d'usage des moteurs n'était pas linéaire : si la part des non utilisateurs et des faibles utilisateurs passe de 24% pour ceux qui se sont connectés avant 1998, à 40% pour ceux qui se sont connectés au deuxième semestre 1999, en revanche, la part des forts utilisateurs est à peu près constante quelle que soit la date de la première connexion, ce qui laisse entendre que le passage au statut de fort utilisateur se fait rapidement pour une frange des nouveaux utilisateurs. On peut faire l'hypothèse que, pour les nouveaux internautes, il existe deux trajectoires assez distinctes : une appropriation rapide des moteurs qui se traduit par un usage intense et à l'inverse un apprentissage difficile qui se traduit par un faible usage voire une absence d'usage[10].



Clef de lecture : parmi les panélistes n'ayant fait aucune requête sur un moteur, 12% se sont connectés à Internet au second semestre 1999

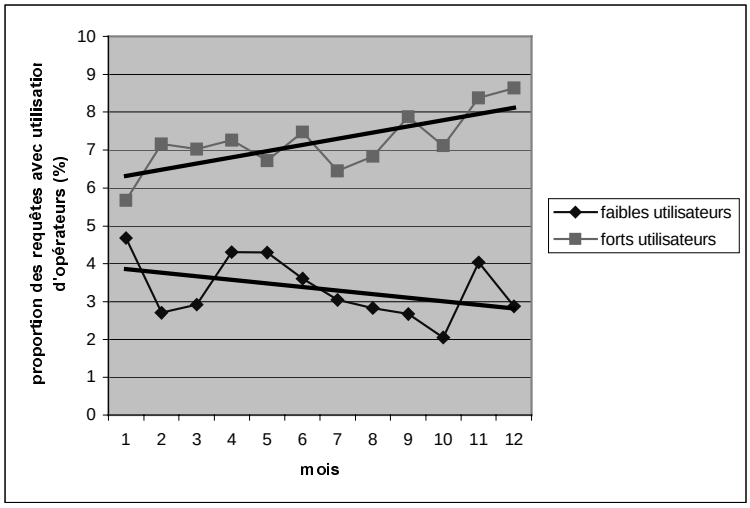
Figure 3 - Ancienneté d'accès à Internet et utilisation des moteurs

La longueur moyenne des requêtes dans nos données est de 1,88 mots (avec un écart-type de 1,22). Ce chiffre est proche de ceux donnés par des études antérieures sur les moteurs de recherche, qui rapportent des moyennes autour de 2 mots-clefs par requête (cf. la revue de ces travaux dans (Jansen & Pooch 2000)). Les trois-quarts des requêtes de notre échantillon ont une longueur inférieure ou égale à 2 mots. L'utilisation des opérateurs est conforme à ce que l'on connaît déjà : la grande majorité des requêtes adressées aux moteurs de recherche (87%) ne contient pas d'opérateurs. 43% des utilisateurs de moteurs de recherche de notre panel n'ont jamais inclus d'opérateurs dans leurs requêtes. Cette proportion de non utilisateurs d'opérateurs varie considérablement selon l'intensité d'usage des moteurs ; en effet, si plus de 80% des faibles utilisateurs (moins de 16 requêtes sur l'année) n'ont jamais utilisé d'opérateurs, cette proportion tombe à moins de 10% parmi les forts utilisateurs (plus de 113 requêtes dans l'année).

La Figure 4 montre l'évolution de l'usage des opérateurs dans les deux populations considérées : nous constatons globalement une nette tendance à la hausse chez les forts utilisateurs et une tendance à la baisse chez les faibles utilisateurs. La courbe concernant la population des faibles utilisateurs est intéressante à observer, si l'on se souvient que cette population est caractérisée par une forte présence d'internautes récents (connectés au deuxième semestre 99, cf. Figure 3 ci-dessus). Nous observons que cette courbe ne « décolle » pas : les faibles utilisateurs, même s'ils ont fait l'effort d'inclure des opérateurs dans certaines de leurs requêtes, ne parviennent pas à acquérir une maîtrise croissante de cette fonction

[10] Cette hypothèse s'inspire de travaux sur l'apprentissage des internautes (Lelong et Thomas, 2001).

avancée. Nous retrouvons probablement ici l'une des deux trajectoires d'usage des moteurs par les nouveaux internautes évoquées ci-dessus, trajectoire qui pourrait conduire à une diminution d'usage, voire à un abandon progressif des moteurs. Les forts utilisateurs, quant à eux, poursuivent une trajectoire d'apprentissage vertueuse : nous observons qu'ils ne cessent d'augmenter leur maîtrise de l'outil en construisant de plus en plus de requêtes complexes.



Clef de lecture : en décembre 2000, environ 9% en moyenne des requêtes moteur émises par les forts utilisateurs de moteurs incluaient au moins un opérateur et cette proportion est de 3% chez les faibles utilisateurs.

Figure 4 – Evolution de la proportion moyenne de requêtes moteur avec utilisation d'opérateurs

Nous avons par ailleurs étudié un autre facteur de complexité des requêtes : le nombre moyen de mots-clefs par requête (longueur moyenne de cette requête). Parmi les faibles utilisateurs, 17% de la population a une longueur moyenne supérieure à 2 mots et cette proportion est de 24% chez les forts utilisateurs. Nous ne notons pas une variation très forte entre les faibles et les forts utilisateurs, surtout si nous comparons les résultats avec la forte variation de l'autre indicateur de complexité (utilisation d'opérateurs) présenté ci-dessus.

Nous pouvons donc conclure que les deux facteurs de complexité d'une requête (sa longueur et l'utilisation d'opérateurs) ne sont pas corrélés de manière identique avec l'intensité d'usage. Le fait d'utiliser des opérateurs dans les requêtes est apparemment un indicateur plus significatif d'un usage avancé.

8. Conclusion et perspectives

Notre étude a concerné un échantillon étendu d'internautes français, dont nous avons suivi les usages d'Internet (et particulièrement ici des moteurs de recherche) sur l'ensemble de l'année 2000. Cet échantillon, qui était en début d'année représentatif de la population connectée à Internet (*cf.* note 5), couvre donc un spectre assez large d'usages et de centres d'intérêts.

Cette étude exploratoire nous a permis d'obtenir des indicateurs quantitatifs précieux et de dresser un portrait des usages des moteurs de recherche par les internautes français. Nous retiendrons que les moteurs de recherche ne sont pas semblables : s'il est déjà notoire que leur recouvrement (en termes de sites indexés) est faible, notre étude a montré que leurs profils d'usage sont nettement différenciés. Les utilisateurs avancés de moteurs profitent apparemment de cette diversité en combinant plusieurs moteurs dans une approche exploratoire ou en sélectionnant, par « affinité thématique », les moteurs qui correspondent au mieux à leurs centres d'intérêt.

L'utilisation des moteurs couvre en réalité des usages et des populations d'internautes très variés. Une exploration fine des parcours dans des contextes d'usage particuliers (achat en ligne, téléchargement gratuit, …), associée à des entretiens et à l'observation de séances de navigation nous permettra d'affiner la signification des faits que nous avons mis en évidence dans cette étude. En somme, un programme de recherche « idéal » traitant de la recherche d'information sur Internet devrait combiner une étude telle que la nôtre, fondée sur les traces d'usages, à une étude telle que (Hölscher & Strube 2000), fondée sur l'observation directe des utilisateurs.

Il nous reste maintenant à situer l'utilisation des moteurs dans un contexte plus global de recherche d'information : quelle est la place du moteur dans un parcours, comment sont appréciés les résultats, comment se poursuit la recherche, comment se décline la stratégie de recherche d'information. Dans cette perspective, les travaux en psychologie cognitive sont d'un grand intérêt, si on les applique à l'observation de situations réelles de recherche d'information.

Références

Amitay E. & Paris C. (2000). Automatically Summarising Web Sites - Is There A Way Around It? *Proc. of CIKM'2000 (ACM 9th International Conference On Information and Knowledge Management)*, pages 173-179.

Benzécri Jean-Paul et coll. (1981). *Pratique de l'analyse des données, Linguistique et lexicologie,* Paris, Dunod.

Bruza, P., R. McArthur, S. Dennis (2000). Interactive Internet search: keyword, directory and query reformulation mechanisms compared. *Proc. of the 23rd Annual International ACM SIGIR Conference*, pages 280-287.

Di Maggio P., Hargittai E., Russell N.W. & Robinson J.P. (2001). Social Implications of the Internet. *Annual Review of Sociology,* 27: 307-336.

Heitzmann R. & Loué J.-F. (2001). L'Internet : les Français se hâtent lentement. *Le 4 pages des statistiques industrielles*, n°152.

Hölscher, C. and G. Strube (2000). Web Search Behavior of Internet Experts and Newbies. *Proc. of the 9th International World Wide Web Conference*.

Jansen, B. J. and U. Pooch (2000). Web user studies: A review and framework for future work. *Journal of the American Society of Information Science and Technology* 52(3): 235-246.

Jansen, B. J., A. Spink, T. Sarasevic (2000). Real Life, Real Users, and Real Needs: A Study and Analysis of User Queries on the Web. *Information Processing and Management* 36(2): 207-227.

Lajoie, J. (2001). Richesse et diversité des requêtes. In J. Guichard ed., *Comprendre les usages de l'Internet*. Paris, Editions ENS.

Lebart Ludovic, Salem André (1994). *Statistique textuelle*. Paris, Dunod.

Lelong B. & Thomas F. (2001). "L'apprentissage de l'internaute : socialisation et autonomisation", *CIUST'01 (Colloque International sur les Usages et les Services des Télécommunications -- e-Usages)*, pages 74-95

Reinert Max (1993). Les "mondes lexicaux" et leur logique. *Langage et société,* 66 : 5-39.

Silverstein, C., M. Henzinger, H. Marais, M. Moricz (1999). Analysis of a Very Large Web Search Engine Query Log. *SIGIR Forum* 33(1): 6-12.