Introduction

En l'espace de quelques années, l'essor du Web s'est accompagné d'une multiplication et d'une diversification de l'offre de contenus, en même temps que d'une généralisation de l'accès aux ressources. La croissance rapide de ce média a suscité dans les premiers temps les spéculations les plus diverses sur des usages encore en construction : « révolution numérique », avènement du « virtuel », nouvelle ère de l'écrit, le Net a été l'objet d'espoirs ou de rejets extrêmes. Depuis lors, de NTICs, *Nouvelles* Technologies de l'Information et de la Communication, en simples TICs, l'objet a perdu l'attrait de la nouveauté tandis que les pratiques tendaient à se banaliser et à se normaliser, et le discours spéculatif a laissé place à l'observation des pratiques réelles. Pour autant, rares sont encore les matériaux d'étude qui permettent d'en rendre compte de manière exhaustive et objective, de sorte que les usages du Web en situation sont encore mal connus et peu décrits : le parcours sur le Web, moment particulier de la rencontre entre production et réception, reste encore à découvrir.

Contexte

En matière de données, pourtant, rarement un support médiatique aura comme le Web donné la possibilité de recueillir autant de traces d'usage : c'est particulièrement le cas du côté des serveurs Web, où l'analyse des *logs*[1] est aujourd'hui monnaie courante. Les visites des internautes y sont enregistrées, comptabilisées et étudiées par les concepteurs de sites pour l'amélioration de leur offre, l'analyse de l'audience, etc. Corrélativement, la connaissance des parcours sur le Web du point de vue des sites est assez avancée aujourd'hui, et constitue un des fondements du *Web Usage Mining*, champ de recherche constitué dans le milieu des années 90 autour de l'analyse des usages du Web. Toutefois, une collection de points de vue centrés sur les sites ne saurait rendre compte des usages individuels : les sites ne connaissent pas plus leurs visiteurs que les contextes d'usage dans lesquels ils s'inscrivent ou les dynamiques personnelles. Malgré cela, les données de trafic centrées-utilisateur sont rares, difficiles à constituer, et seule une poignée de travaux ont pu disposer d'un tel matériau, pour des études de courte durée.

D'autres approches plus qualitatives ont su se placer du côté de l'utilisateur pour en observer finement les pratiques. Les sciences cognitives, dans la lignée des travaux sur les hypermédias, ont pratiqué des expériences en laboratoire afin d'étudier des comportements dans des contextes précis, notamment la recherche d'information ; les conclusions de ces travaux viennent le plus souvent alimenter un projet plus

[1] Historique de l'ensemble des requêtes adressées à un serveur ; voir Glossaire.

global de modélisation de l'utilisateur et de recherche de modèles mentaux impliqués lors de la navigation sur le Web. Malgré des tentatives pour élaborer des modèles cognitifs capables d'expliquer l'ensemble des comportements sur le Web, ces dispositifs se heurtent à un problème de généralisation des résultats à partir des expérimentations locales, et peinent à rendre compte de la diversité des situations rencontrées par les internautes.

Dans un autre champ disciplinaire, la sociologie des usages, l'ethnométhodologie et les sciences de l'information et de la communication ont cherché à décrire les pratiques en situation naturelle, par le biais de questionnaires, d'entretiens, d'observations ou d'enregistrements vidéo. Si ces approches ont su mettre à jour les implications sociales, interactionnelles et sémiotiques des TICS, elles peinent, sur la question des parcours, à atteindre une analyse à la fois fine et globale. D'une part, la méthodologie des questionnaires ou des entretiens, au-delà du statut particulier que l'on peut conférer aux déclarations et aux discours tenus dans ce cadre, dressent des descriptions à gros grain des pratiques ; d'autre part, les méthodes d'observation directe ont pour elles la richesse d'une description détaillée des modes d'activité, mais elles peinent à rendre compte de la globalité des pratiques et des situations et se heurtent à un problème de « masse critique » des données.

Une description tout à la fois globale et fine des usages de la Toile reste donc à construire, sous la forme d'une typologie des comportements de navigation tenant compte de l'offre de contenus du Web, du contexte local de l'utilisateur et de la dynamique de ses pratiques personnelles, et de la construction globale de pratiques normées par l'ensemble des internautes. Elle s'appuiera à bon droit sur des données de trafic centrées-utilisateur, qui offrent tout à la fois une perception exacte des contenus visités et une vue dans la durée de la diversité des situations d'usage. Pour cela, elle nécessite d'élaborer des méthodes et des outils capables de tenir compte de la spécificité des contenus Web et de leur mode d'appréhension, afin de rendre possible l'élaboration d'une sémantique des parcours.

Objet d'analyse et champs disciplinaires

Cette étude entend apporter une contribution à la description des usages du Web, fondée sur l'analyse de corpus de parcours effectués en situation naturelle. Nous prenons pour cela appui des données de trafic d'internautes recueillies auprès de trois panels : un panel représentatif de plus de 3 300 internautes en 2002, une cohorte de 600 personnes observées sur trois ans, et un panel restreint d'utilisateurs des bibliothèques numériques. Ce travail, mené au sein du laboratoire Usages, Créativité, Ergonomie de France Télécom R&D, s'inscrit dans le cadre plus large de projets d'étude des usages d'Internet à domicile[1].

L'objet d'étude est le parcours, c'est-à-dire la visite ordonnée et déterminée temporellement de pages et de sites Web, dans le cadre d'une unité d'action cohérente, la session. L'analyse des parcours se fait en corpus : de la même manière que la linguistique de corpus a profondément renouvelé l'approche du matériau

[1] Projets TypWeb, SensNet et BibUsages ; voir Annexe 1 pour une description complète.

textuel en s'appuyant notamment sur l'analyse de masses de documents numérisés, l'étude des parcours sur le Web s'appuie ici sur un vaste corpus de parcours au sein duquel nous souhaitons distinguer des comportements significatifs et récurrents.

Pour manipuler cet objet particulier, notre travail mobilise plusieurs disciplines :

- *sémantique textuelle* : la sémantique interprétative textuelle a montré, du palier des lexèmes à celui des textes, la construction contextuelle du sens, la détermination du local par le global et l'inscription des pratiques d'écriture et de lecture dans des genres et des situations. Les contenus du Web et leur appréhension n'échappent pas à ces déterminations sémantiques ; ils ne sauraient toutefois s'y réduire, pour deux raisons. D'une part, la dimension hypertextuelle du média tend à briser les unités textuelles en privilégiant le fragment et la recomposition d'un ensemble à partir de sources au sein du parcours ; d'autre part, les contenus Web ne peuvent être envisagés sous l'angle des textes uniquement, mais également comme un espace d'action et d'outils mis à disposition de l'internaute. Une sémantique des parcours s'appuiera donc sur une sémantique textuelle adaptée aux spécificités des contenus du Web et de leurs modes d'appréhension.
- *sociologie des usages* : l'analyse des parcours tire parti des travaux déjà menés dans le champ de la sociologie et de l'action située. D'un côté, les enjeux de l'usage des TICS sous l'angle des inégalités (thème du « fossé numérique »), de la constitution des communautés et des collectifs, et de l'impact sur les organisations forment un entour global qui guide l'interprétation des modes de navigation ; de l'autre, les approches plus praxéologiques centrées sur l'analyse de l'action en situation, notamment dans le champ de l'ethnométhodologie, apportent des descriptions situées des usages auxquelles se rattache plus particulièrement notre travail.
- *analyse de données et Web Mining* : l'analyse de données de trafic volumineuses et centrées-utilisateur nécessite de se doter des outils nécessaires à leur manipulation. Il est donc question d'informatique, à double titre : en premier lieu, le substrat technique des données et du Web lui-même doit être connu pour être manipulé ; ensuite, le caractère exploratoire de cette recherche implique de tester les limites des outils, des méthodes statistiques et des représentations existants, et d'être capable d'en produire de nouveaux pour manipuler les corpus de parcours.

Et, au-delà de l'inscription pluridisciplinaire de l'analyse de corpus de parcours, notre travail se place résolument dans le champ des sciences humaines, sous l'angle de la description des pratiques.

Notre contribution

Cette étude est exploratoire à double titre : d'une part, elle vise une description des pratiques de navigation à domicile sur des durées et des panels inédits à ce niveau de détail. D'autre part, elle s'appuie sur des données de trafic centrées-utilisateur, dont l'exploitation encore rare nécessite la mise en place d'outils et de méthodes *ad hoc*.

1. *Méthodologie et outils pour analyser les données de trafic centrées-utilisateur*
   L'analyse de traces de navigation recueillies sur les serveurs Web propose des méthodes statistiques qui ne peuvent être appliquées aux données centrées-utilisateur pour deux raisons majeures : elles laissent de côté la question des contenus (connus, lorsque la navigation est analysée sous l'angle des sites), et elles reposent sur une redondance dans les données que l'on n'observe pas sitôt qu'on se place du côté de l'utilisateur. La diversité des contenus, des modes de navigation, des formes de parcours, nécessite la mise en place de méthodes d'analyse originales qui représentent tout à la fois les contenus visités et la dynamique des parcours. Sur le plan du contenu, notre travail met en œuvre et évalue différentes stratégies pour attacher aux listes d'adresses des pages visitées (ou *URL*[1]) par les internautes des informations sur les thématiques et les fonctions des pages et des sites vus. Sur le plan de la navigation dans la session, nous élaborons des indicateurs statistiques capables de représenter la forme et la rythmique des parcours.
   Le parcours devient alors un objet complexe à analyser, hétérogène du fait de ses constituants et de ses différentes métriques. Pour l'aborder, nous mettons en place des outils de fouille manuelle des données de trafic qui permettent d'approcher au plus près la réalité des parcours et leur logique, et nous proposons une démarche statistique descriptive qui tient compte de ces particularités.

2. *Segmentation des parcours et description des pratiques de navigation*
   Les méthodes et outils que nous mettons en place pour décrire les parcours sur le Web ne sont valables que tant qu'il servent effectivement l'analyse et la caractérisation des comportements de navigation. Les corpus de parcours dont nous disposons, inédits par leur taille et leur durée d'observation, nous permettent de construire une segmentation des parcours fondée sur l'observation de régularités et la mise à jour de contextes d'usages prototypiques. Les données longitudinales exhaustives permettent également d'observer la structure et l'évolution des territoires personnels sur le Web ; confrontée aux modes prototypiques de navigation, cette vision éthologique des parcours nous amène à distinguer des modes d'appréhension type des contenus dans le contexte de l'usage.

Organisation du mémoire

Cette thèse se décompose en deux grandes étapes, qui renvoient au caractère doublement exploratoire de notre travail. La première partie s'attache à décrire les méthodes et outils que nous avons été amenés à élaborer pour analyser les données des trafic centrées-utilisateur ; la seconde mobilise cet outillage pour l'étude des pratiques de navigation, et en propose une segmentation fine sous l'angle de la forme des parcours, de leur contenu et des territoires personnels sur le Web.

---

[1] Voir Glossaire.

Nous posons dans le Chapitre 1 un cadre d'analyse et de réflexion pour appréhender les parcours sur le Web, et positionnons notre travail par rapport aux études déjà menées sur cet objet. Après une présentation du type de données dont on dispose, le Chapitre 2 expose les différentes étapes de leur mise en forme : il s'agit de les nettoyer des scories qu'elles contiennent, et d'identifier, au sein des listes horodatées d'URL, des sessions, des pages, des sites. Le Chapitre 3 est consacré aux différentes stratégies mises en œuvre pour attacher aux données de trafic des informations de contenu : exploitation des URL brutes, aspiration de pages, catégorisation semi-automatique des URL, mobilisation des annuaires du Web. Cette caractérisation des contenus est complétée sur le plan de la dynamique des parcours par l'élaboration d'outils de fouille manuelle et d'indicateurs statistiques représentant la forme et la rythmique des sessions, qui sont exposés au Chapitre 4.

La seconde partie montre l'exploitation de cet outillage pour la description des usages du Web et la segmentation des parcours en situation. Le Chapitre 5 décrit tout d'abord la composition des trois panels sur lesquels se base cette étude ; il propose ensuite deux explorations des données sur la base de leur forme d'une part et de leur contenu de l'autre, qui permettent d'en dresser le profil et le comportement statistique. Ces éléments servent de base à une classification des sessions en cinq groupes à partir des indicateurs topologiques, sur lesquels nous projetons les contenus visités pour apercevoir le lien fort qui unit ces deux composantes. Le Chapitre 6 confronte ces profils-type de parcours avec des éléments de contexte relatifs à l'utilisateur : en s'appuyant sur la structure et la dynamique des territoires personnels sur le Web, nous mettons à jour trois modes d'appréhension prototypiques du Web qui impliquent tout à la fois les types de contenus et la dynamique de l'usage. L'examen spécifique des modes d'accès et de manipulation des bibliothèques électroniques et des fonds numérisés permet d'approfondir ce problème, et montre la prévalence de l'usage sur la nature propre des documents lorsque ceux-ci sont immergés dans le contexte du Web.