Thèse de doctorat de l'Université de Paris X – Nanterre

Spécialité : Sciences du langage

présentée par
Thomas BEAUVISAGE

pour obtenir le grade de Docteur de l'Université de Paris X

Sémantique des parcours des utilisateurs sur le Web

Sous la direction de François RASTIER

Soutenue en octobre 2004

Devant le jury composé de :

M. Houssem ASSADI
M. Dominique BOULLIER (rapporteur)
M. Benoît HABERT
M. Ludovic LEBART
M. François RASTIER (directeur)
M. Pierre ZWEIGENBAUM (rapporteur)

Résumé

Notre thèse a pour objectif de décrire les parcours sur le Web sur la base de données de trafic centrées-utilisateur. Nous proposons des méthodes et des outils pour enrichir de telles données de trafic, et les mettons en application pour construire une segmentation des parcours sur la base de leur forme, de leur temporalité, de leur contenu et de leur insertion dans les pratiques individuelles. Ce travail, mené au laboratoire Usages, Créativité, Ergonomie de France Télécom R&D, s'inscrit dans le projet SensNet qui vise à analyser les usages d'Internet à domicile.

La généralisation de l'accès à Internet en France entraîne une banalisation et une normalisation des pratiques du Web. Pour autant, l'activité de navigation reste mal connue : si l'analyse des *logs* des serveurs Web est maintenant bien maîtrisée, celle des traces de navigation recueillies du côté de l'internaute en situation naturelle demeure rare et complexe. Les données utilisées dans cette étude centrée-utilisateur proviennent de sondes de recueil de trafic Internet installées sur les postes des utilisateurs à domicile ; on obtient alors la liste des URL visitées par chaque internaute, qui constitue le matériau premier de l'étude. Sur cette base, nous proposons une description des parcours des internautes de page en page et de site en site centrée sur la session. Cette description intègre les informations sur les contenus visités d'une part et les territoires personnels sur le Web d'autre part, et examine leur articulation dynamique au sein des parcours.

Pour y parvenir, un premier travail consiste, après une première mise en forme de ces données brutes, à les enrichir. Sur le plan des contenus, nous proposons une méthode qui exploite les informations fournies par les annuaires du Web pour qualifier les URL visitées. Adossée à un module d'identification des services sur les portails généralistes développé dans le cadre du projet SensNet, cette description permet d'appréhender l'offre de contenus du Web dans sa diversité : informations, mais aussi services, outils, fonctionnalités. Sur le plan de la navigation, nous élaborons des indicateurs statistiques simples qui rendent compte de la forme, de la temporalité et du rythme des parcours, à l'échelle de la page et du site. En complément de cette approche *macro*, nous avons développé des outils de fouille manuelle des sessions permettant de vérifier les résultats de l'approche quantitative et de formuler des hypothèses sur les comportements des internautes. Ainsi dotés, nous disposons des outils nécessaires pour observer, au sein de données volumineuses, les liens entre forme et contenus des parcours, et mettre à jour des régularités dans les pratiques des internautes.

Nous appliquons cet outillage à 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. Ces trois sources de données complémentaires nous amènent à établir une première typologie des sessions sur la base de leur forme et de leur temporalité : les cinq parcours-type mis à

jour s'opposent sur le plan de leur durée, de leur forme et de leur rythmique, et montrent la grande diversité des comportements.

Examinés sous l'angle des territoires personnels, ces modes prototypiques de navigation prennent sens. Au sein d'espaces Web *a priori* non bornés, les internautes dessinent des zones familières de taille restreinte autour de thématiques propres à chacun. Trois zones distinctes sont mises en évidence, auxquelles correspondent des modes d'activité et des types de contenus spécifiques : le familier, orienté vers des contenus à fort taux de renouvellement (flux d'information, services de communication), constitue le noyau dur de l'activité de navigation, et induit des parcours routiniers rapides et ciblés qui s'apparentent aux modes de consommation des média traditionnels (télévision, radio, journaux). Le territoire occasionnel délimite des zones visitées moins fréquemment, mais de manière régulière dans un contexte donné, et cible les contenus de type service ou achat : dans ce cadre, les sessions s'allongent et se complexifient, mais l'espace hypertextuel demeure connu et maîtrisé. Enfin, les parcours de découverte amènent l'internaute à mobiliser le Web comme ressource informationnelle ponctuelle de manière ciblée : dans ces sessions où la ligne brisée domine, les moteurs de recherche dessinent un espace de sites que l'utilisateur ne reverra plus pour la majorité d'entre eux.

Sur le plan méthodologique, ces résultats attestent la capacité de notre outillage à décrire et expliquer les comportements de navigation sur le Web ; ils montrent également la nécessité pour une sémantique des parcours de tenir compte des déterminations globales pour comprendre les comportements locaux, et de mener l'étude des usages sous une angle praxéologique.

Sur le plan des pratiques, on observe ainsi que le parcours Web est la résultante d'une double dynamique, celle des contenus proposés et celle de l'utilisateur, dont la confrontation induit des modalités d'activité qui dépendent autant des contenus eux-mêmes que de leur appréhension et de leur valorisation par l'utilisateur. Loin de « surfer » au gré des hyperliens, l'internaute construit, au sein d'un vaste espace hypertextuel, des zones restreintes de familiarité qui constituent l'essentiel de ses pratiques sur le Web.

Abstract

This thesis aims at describing users' paths through the Web on the basis of user-centric traffic data. We propose methods and tools to enrich traffic data, and apply them to build a segmentation of Web paths based on their shape, their temporality, their content and their place in individual practices. Our work took place in the Uses, Creativity, Ergonomics laboratory at France Telecom R&D, within a project named SensNet dealing with the analysis of domestic uses of the Web.

The generalization of Internet access in France leads to a normalization of Web practices. However, the activity of Web browsing itself remains rather unknown: while the analysis of Web servers access logs is now widely practiced, those of user-centric real-world traffic data is still rare and complex. This study relies on the analysis of data collected by probes installed on users' computers at home, which provide the time-stamped list of all the URLs visited by each Internet user. On this basis, we propose a description of Web users' paths through pages and sites centred on the session. This description integrates information on the content of pages and sites as well as on personal territories on the Web, and examines their dynamic articulation inside Web paths.

To achieve this goal, after data preparation for the analysis, we have to enrich them first. On the side of content description, we propose a method which exploits information provided by Web directories to qualify the visited URLs. Combined with a module for identifying services on generalist portals developed within the SensNet project, this description reflects the diversity of Web contents: information, but also services, tools, functionalities. On the side of browsing, we calculate robust statistical indicators which represent the form, the temporality and the rhythm of Web paths, both at page-scale and site-scale. Beside this *macro* approach, we developed tools for manually exploring sessions, that allow to verify the results of quantitative approach and to formulate hypothesis concerning Internet users' behaviour. Thus, we have the necessary tools to observe inside large datasets, links between paths' topology and content, and to highlight regularities within Web users' practices.

We apply these tools to three panels: a representative panel of more than 3.300 users in 2002, a cohort of 600 people observed during three years, and a small panel of digital libraries users. These three complementary datasets allow us to build a typology of sessions based on their topology and their temporality: the five discovered types of paths differ in terms of duration, form and rhythm, and demonstrate the great diversity of browsing behaviours.

These prototypical modes of navigation make sense when considered from the angle of personal Web territories. Within *a priori* unlimited spaces, Web users outline small zones related to specific topics. Three distinct zones are identified, which correspond to particular modes of activity and content types: the familiar

territory, oriented on regularly updated contents (information streams and communication services), forms the core of users' browsing, and implicates fast and targeted routine paths related to traditional mass media consuming modes (television, radio, newspapers). The occasional territory refers to zones which are less often visited, but regularly in a given context, and to service and e-commerce contents: in that case, Web paths are longer and more complex, whereas the hypertextual space still remains well-known and under control. Finally, in discovery paths, Internet users make use of the Web as information resource for targeted searches: in these highly non-linear sessions, search engines are often mobilized to explore Web spaces which will, for most of them, never be visited again by the user.

On the methodological side, these results attest the ability of our tools to describe and explain navigation behaviours on the Web; they also demonstrate the necessity for a semantics of Web paths to take into account global factors to understand local behaviours, and to have a praxeological approach of usage studies.

On the side of practices, we observe that a Web path results from a two dynamics: the one of the proposed contents by Web sites, and the one of the user. Their confrontation in context implicate distinct modes of activity which depend as much on the visited contents as on their reception and their valuation by the user. Far from wildly "surfing" the Internet from link to link, Web users define, within a vast hypertextual space, restricted familiar zones that constitute the core of their practices on the Web.

Remerciements

Ce travail de thèse m'aura au moins appris deux choses. La première est que la recherche est une affaire de collaborations plus que d'individualités, et qu'aucun travail sérieux ne saurait être mené sans être inscrit, de quelque manière et à quelque niveau que ce soit, au sein d'un collectif de recherche.

Je ne saurai donc manquer de remercier toutes les personnes avec lesquelles j'ai été amené à collaborer au cours de ce travail : François Rastier, en premier lieu, m'a témoigné sa confiance en acceptant de diriger cette thèse qui se nourrit abondamment de son travail et de ses conseils ; Valérie Beaudouin m'a chaleureusement accueilli au sein du laboratoire Usages, Créativité, Ergonomie de France Télécom R&D, et Houssem Assadi a encadré et guidé ce travail : tous deux m'ont témoigné une disponibilité et un soutien sans faille ; et les participants, réguliers ou occasionnels, des projets TypWeb et SensNet : la mise en commun des résultats, les réunions régulières et les discussions critiques en ont fait un projet efficace, vivant et fondamentalement collectif sans lequel notre travail n'aurait pu aboutir.

En second lieu, cette thèse m'a montré qu'un bon lecteur est une chose précieuse. Outre ceux qui ont accompagné ce travail, je tiens à remercier Dominique Boullier, Benoît Habert, Ludovic Lebart et Pierre Zweigenbaum d'avoir accepté de faire partie du jury de cette thèse et d'en être les lecteurs privilégiés. C'est avec plaisir que je soumets ce mémoire à leur jugement et leurs critiques.

Enfin, j'ai pu vérifier au cours de ces quelques années de travail que les discussions de couloir sont au moins aussi importantes que les échanges plus formels, et que ces à-côté prennent parfois la forme de l'essentiel. Merci donc à Thomas, Julien, Marie, Julia, Marc, Shark et les autres, qui ont contribué à ce travail bien plus qu'ils ne le croient.

Sommaire