De l’analyse de traces à l'exploitation des phénomènes d'intelligence collective

Luigi Lancieri

France Telecom R&D
42 rue des coutures
14000 Caen

luigi.lancieri@orange-ftgroup.com

RÉSUMÉ. *Le dénominateur commun de toutes les formes de traces est le statut de mémoire qui leur est associé. Ce constat implique qu'il est possible de puiser a posteriori dans cette mémoire et d'en extraire de la « connaissance » liée à l’activité d’un individu ou d’un système. L'informatique a produit deux changements importants dans cette réflexion : le caractère numérique, stockable et manipulable des traces et la croissance de la part des activités humaines impliquant d’une manière ou d’une autre un ordinateur. Compte tenu du fait que les individus interagissent de plus en plus par média interposé (création de contenu numérique, échange de courrier électronique, achat en ligne, etc), les données qu'ils produisent peuvent faire l'objet d'un traitement rapide et automatique. Dans ce contexte l’exploitation de traces d’activités permet de mieux comprendre le comportement des individus et d’utiliser la mémoire de leur contexte d’activité pour offrir de nouveaux services.*

Nous proposons de faire une synthèse de ces questions dans le domaine des interactions humaines médiatisées. Après avoir posé le cadre de notre réflexion, nous décrivons des aspects méthodologiques et des éléments de modèles en insistant sur les conclusions de type sociocognitives dérivées de l'analyse de données quantitatives. Dans un second temps, nous donnons quelques exemples d’utilisation concrète de données issues de comportements collectifs pour rendre des services.

MOTS-CLÉS : *Comportements collectifs, mobilité, analyse de traces, systèmes de recommandation*

KEYWORDS: *Collective behaviour, mobility, traces analysis, recommendation systems*

Analyse de traces et personnalisation des environnements informatiques pour l' apprentissage humain (Hermes Science)

1. Introduction

Nous abordons dans ce chapitre l'étude des phénomènes collectifs sous un angle métrologique et systémique au sens de la vision cybernétique proposée par N.Wiener (Wiener, 1952). Selon notre approche un groupe est un ensemble, plus ou moins, cohérent d'individus en interaction. La notion de cohérence que nous développerons un peu plus loin implique un certain niveau de proximité dans les objectifs, les centres d'intérêts ou l'activité des membres du groupe. Il s'agit d'une contrainte forte pour une approche de type métrologique qui pourrait être perçue de manière purement quantitative et s'accommoder de traitements statistiques. Dans la pratique cette contrainte peut être facilement satisfaite si on considère que la plupart du temps, les individus se rendant dans un même lieu, virtuel ou physique ont des intérêts ou des activités proches. Nous partons donc de l'hypothèse qu'en termes de probabilité le regroupement a un sens. Cette hypothèse nous permet de donner à nos mesures, par essence quantitatives, une interprétation qualitative. Notre approche n'est donc pas spécifique aux EIAH et peut être facilement transposable à d'autres contextes où la notion de groupe est pertinente. Nous pensons cependant que par nature le monde de la formation, en considérant de près les phénomènes collectifs de création, de partage de la connaissance et de co-construction, est sans aucun doute un environnement d'étude privilégié.

L'analyse des traces n'est pas un exercice trivial. Au delà de difficultés scientifiques liées à l'interprétation et la validation des résultats se pose aussi le problème technique de la collecte des données dans le respect des contraintes éthiques et juridiques. Ceci étant dit, l'enjeu est important dans la mesure où il est envisageable d'obtenir une qualité avancée de l'observation compte tenu de la grande quantité de données disponibles. Sous cet angle l'analyse de traces peut être un complément utile de l'observation psychosociale dans la compréhension des phénomènes collectifs. Le second enjeu est la réinjection de la connaissance contextuelle extraite des données dans des services à valeur ajoutée. Les cas du filtrage collaboratif et des systèmes de recommandation sont de bons exemples d'utilisation utiles des traces d'activités.

Avant de commencer à débattre de ces questions nous proposons quelques définitions ainsi que des modèles d'interactions médiatisées dans lesquels le media et l'environnement jouent un rôle actif. Nous abordons ensuite la question de l'interprétation des traces en proposant quelques méthodes pour prendre en compte la structure des interactions. Enfin, avant de conclure par une analyse de services bénéficiant des traces d'interactions, nous posons quelques éléments de réflexion sur les questions éthiques et juridiques.

2. Média et environnement, plus que des supports

A priori, le rapport entre le média et l'environnement peut paraître très lointain. Nous verrons dans cette section que ces deux notions peuvent se fondre dans le concept de média généralisé, particulièrement adapté dans des formes de communications pervasives (Derycke et al, 2007, Ottavi et al 2007).

2.1. Le média

Selon les communautés le terme média peut avoir une définition différente. Dans un sens strict, on nomme média un moyen impersonnel de diffusion d'informations, utilisé pour communiquer en principe, sans possibilité de personnalisation. Dans une vision plus ouverte le media a le sens d'intermédiaire et regroupe les composants (matériel, logiciel, contenu) permettant l'interaction entre les individus. Les outils associés à Internet comme la messagerie ou les serveurs web sont donc des média et supportent naturellement les traces d'interactions. Sous cet angle, le média est impliqué dans une influence mutuelle et complexe avec l'interaction.

D'une part, le média influence l’interaction dans la mesure où il modifie la perception de la réalité. Citons l'exemple de l'ergonomie des interfaces, dont la capacité à transmettre des formes plus ou moins réduites d'un message peut en orienter ou influencer l'usage. Il est notoire par exemple que le web, le mail ou le téléphone mobile ont modifié les habitudes d'échanges entre individus. Les sociologues ont aussi montré que les relations de confiance ou les modes de résolution des conflits étaient fortement dépendants du media. Ainsi, plus le niveau de réduction opéré par le média diminue (e.g. chat, téléphone, face à face) plus les relations de confiance s'établissent favorablement et contribuent à la résolution des conflits. Dans le cas inverse, les relations entre les individus restent la plupart du temps dans l'état précédant l'interaction (Doise, 2003).

D'autre part, et même si cela parait moins évident à première vue, la relation réciproque est vraie dans la mesure où l'interaction a aussi une influence sur les fonctionnalités et les performances du media. Le confort de navigation, la latence dans le réseau, en un mot certains attributs du média sont influencés par l’activité des usagers. Dans les services d'échanges de fichiers (e.g. P2P), la popularité des documents échangés (i.e les goûts, les modes,..) a une influence sur la rapidité (confort) d'accès à l'information. On peut fournir de nombreux exemples allant dans le même sens : les systèmes de recommandation d'achat en ligne, la notation d'expertise dans les forums de discussions sont autant de cas où l'activité des individus va modifier la performance du média. Dans le cas de la recherche d'information, les technologies actuelles comme le page rank ont tendance à modifier

le classement des réponses à une requête en fonction des interactions (e.g goûts, modes, liens vers des sites populaires, boycott des sites concurrents, ...)

Une caractéristique remarquable du média, selon notre perspective d'analyse, c'est qu'il mémorise l'interaction. La convergence des services vers les technologies Internet va dans le sens de la généralisation de cette remarque. Même des services comme la radio ou la télévision (i.e. sur Internet) suivent cette tendance avec des performances, une profusion de contenus, une ergonomie et des capacités de personnalisation qui peuvent rapidement faire la différence avec le service originel. Tous ces médias basés sur les ordinateurs sont fournisseurs d'une grande quantité de traces qui correspondent à une mémorisation de l'activité des usagers mais aussi de leur contexte. Cette notion est importante car la prise en compte du contexte est une différence clé pour distinguer l'information de la connaissance, qui est la véritable plus-value pour une réutilisation des traces d'activité (voir section 5).

2.2. L'environnement

La relation entre le média et le contexte est liée à l'usage que fait un individu du média à un moment ou dans un lieu donné et avec des objectifs également donnés. Si on considère que le temps et l'espace ont une influence sur l'usage, par le biais par exemple de lieux ou de moments préférentiels d'exercice de l'activité alors il convient d'associer ces dimensions au contexte. En poussant cette logique aux limites et dans une perspective d'activité pervasive, l'environnement devient une partie à part entière du média. Dans le pervasive computing (intelligence ambiante), les capacités de traitement, de stockage et de communication sont granularisées et réparties à l'extrême (vêtements, bornes, objets actifs, domotique, infrastructure..). L'environnement acquiert alors les fonctionnalités du média.

Pour clarifier les idées, imaginons un usager en vacances qui recevrait spontanément sur son mobile un SMS lui indiquant que le monument qu'il a en face de lui a une histoire hors du commun. Le moment et le lieu de la transmission de cette information ont un sens. Dans ce cas de figure, non seulement la vision du contexte est étendue mais en plus, la mémoire de l'activité et du contexte est prise en charge par des fonctions fondues dans l'environnement. Cet exemple n'est pas complètement futuriste. La ville de Strasbourg a en effet récemment ouvert un nouveau musée remarquable où, grâce à des capteurs de présences, le visiteur muni d'un terminal reçoit tout au long de sa visite des commentaires audio directement et très précisément en rapport avec la zone de l'espace du musée où il se trouve. La technologie permet aujourd'hui d'envisager ces services au-delà de zones confinées, services qui peuvent être étendus aux villes ou à des espaces plus vastes du territoire.

Avec un statut de média, l'environnement hérite aussi ses propriétés et en particulier l'influence mutuelle avec l'interaction. En effet, non seulement l'information que me transmet mon mobile est dépendante du lieu où je me trouve ou des individus que je rencontre mais réciproquement, les capacités informationnelles (communications, ...)

de l'environnement peuvent conditionner les endroits où je me rends. Ce principe est assez classique et fait que les individus ont déjà tendance à se rendre en des lieux où sont déjà massivement présents d'autres personnes (voir plus loin). Au départ, les individus choisissent des endroits et des moments en fonction de leurs caractéristiques propres (proximité, agréable, fonctionnel, ...) et parce que d'autres individus les ont choisis. Ces choix modifient les contraintes d'espaces et de temps (agencement de l'espace, places disponibles, ...) qui rétroagissent sur le choix des individus. Au travers de ce jeu d'influence mutuelle, environnements et médias deviennent adaptables et actifs. Ils produisent des traces et jouent un rôle fusionnel avec les individus.

2.3. Modèles d'interactions médiatisées

Dans le modèle d'interaction médiatisée que nous proposons, l'interaction résulte de la fusion de différents éléments de base.

- Organisation : un vers un, un vers plusieurs, etc.
- Le temps : niveau d'asynchronisme, durée de l'interaction, etc.
- Espace d'échange : profondeur de la mémoire partagée, niveau de visibilité des actions individuelles, niveau de fusion entre espace physique et informationnel, etc.

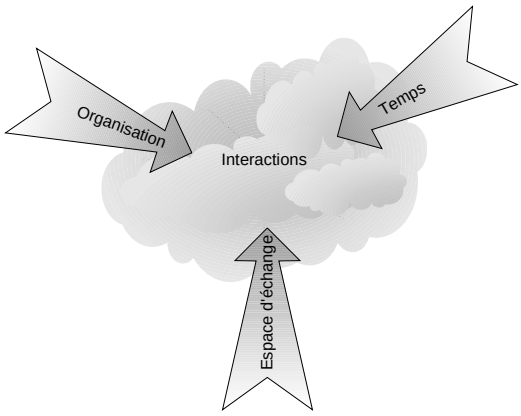

Figure 1: Fusion des composantes de l'interaction

Ce modèle est dynamique dans la mesure où chaque contributeur dans l'interaction, qu'il s'agisse d'un humain ou d'une machine incluant le média généralisé, a une vision de l'interaction et a la capacité à s'y adapter. Cette symbiose homme-machine est largement dépendante des traces que peuvent produire les différents acteurs de l'interaction et qui constituent un lien de rétroaction dans une perspective systémique.

3. Analyse de traces, des données au sens

Les traces sont donc une mémoire de la combinaison complexe des interactions entre les individus et entre individus et machines. Nous proposons une vision élargie de cette notion qui recouvre plutôt d'habitude une simple suite d'éléments d'information de bas niveau.

Dans notre perspective, les traces correspondent à toute forme d'information produite consécutivement à l'activité d'un individu ou d'un système. Le fichier de log d'un serveur comme le contenu d'un mail sont donc autant de traces et à ce titre ont un statut de mémoire et témoignent de l'activité. Par ailleurs, la production de traces s'inscrit dans une démarche d'écologie informationnelle, ce qui implique que l'analyse de traces ou l'exploitation des résultats associés n'est pas une fin en soi (Lancieri, 2001). L'activité, la liberté de choix de l'individu est structurante et doit dans tous les cas garder la priorité. La démarche écologique est avant tout une posture philosophique qui garantit un garde-fou minimal sur le plan éthique (voir plus loin), mais elle a aussi un intérêt en termes de productivité et de représentativité des modèles. Le fait de n'exploiter que les traces naturelles existantes évite d'abord les surcoûts liés à des architectures dédiées. Par ailleurs, le suivi forcé de l'activité des usagers peut engendrer des effets de bord bien connus en métrologie où l'acte de mesure modifie la réalité mesurée. Dans cette vision écologique, les intérêts, éthiques, scientifiques et économiques se combinent mutuellement.

Pratiquement, les traces sont utilisées après extraction d'une combinaison plus ou moins complexe de descripteurs. Au delà des grandeurs purement quantitatives, comme le nombre d'usagers accédant à telle ou telle autre ressource, l'utilisation des mots-clés ou d'autres éléments relatifs à la sémantique des contenus est souvent plus riche pour modéliser l'activité et permet une meilleure caractérisation de l'accès à la connaissance. Comme nous le verrons plus loin des éléments liés au déplacement ou à la chronologie de l'activité peuvent aussi être intégrés à ces descripteurs. Notons au passage que des éléments quantitatifs peuvent être interprétés d'un point de vue qualitatif et révéler l'intérêt pour certaines formes d'activité ou certains contenus. Les descripteurs peuvent être réunis au sein d'une structure formelle (e.g un vecteur, structure XML, ...) et être utilisés dans des mesures de distances qui pourront mettre en évidence des similarités. On considère dans cette approche que l'ensemble des contenus consultés sur une période fournit des éléments représentatifs du profil de l'usager. Il sera aussi possible d'utiliser des techniques de fouille de données pour mettre en évidence des régularités et identifier des groupes d'intérêts ou de pratiques.

Ceci étant dit, une question essentielle, de notre point de vue, concerne la représentativité des structures de connaissances construites à partir des traces. Autrement dit, et si on considère l'exemple des vecteurs de mots clés, deux vecteurs proches impliquent-ils réellement la similarité dans l'activité des deux usagers concernés ? Les travaux que nous avons menés confirment l'intuition, en tout cas dans le contexte de notre étude. On observe en effet une corrélation de 70% entre les

similitudes de consultation de site web et la proximité des profils de mots clés (Lancieri, 2005).

Quoique réductrice, la mise en œuvre de métriques applicables à des éléments intangibles comme le profil d'un usager jette un pont entre le modèle et les applications concrètes. Il est cependant indispensable d'évaluer, au préalable, la représentativité de ces profils, ce qui est rarement fait dans la pratique.

Au-delà de l'activité individuelle, cette philosophie peut être suivie pour mieux comprendre les comportements collectifs qui sont plus difficiles à appréhender par une observation psychosociale. En effet, l'activité collective met en œuvre des mécanismes d'influences et de synergies empreints de complexités qui dépassent parfois la conscience des individus. Ces mécanismes engendrent des effets comparables à une forme d'énergie qui se manifeste par l'activité du groupe et que l'on peut mesurer et envisager de réutiliser. La ré-exploitation de l'intelligence collective latente est un autre aspect de l'écologie informationnelle.

3.1. Traces de consultations et activité collective.

L'exemple qui suit va nous permettre d'illustrer comment l'analyse de traces d'activités, même de relativement bas niveau, peut être utile pour mieux comprendre les comportements collectifs. En observant d'un point de vue quantitatif les consultations des usagers sur les sites web nous pouvons faire différentes remarques susceptibles de s'interpréter d'un point de vue qualitatif.

En calculant globalement la distribution statistique du nombre d'hyperliens par pages web on observe que la plupart des pages contiennent entre 8 et 68 liens. On peut déterminer pour une page donnée une caractéristique plus significative qui est la densité d'hyperliens (nombre de liens rapportés au nombre de mots clés non vides contenus dans la page). Cette valeur peut aussi être analysée en observant que le nombre d'hyperliens par page est un indice de la connexité du web, et que l'évaluation de cette donnée dans le temps peut être un indicateur de l'évolution du web en tant qu'écosystème informationnel (i.e. collectif). Cet indicateur permet aussi d'évaluer une dimension particulière du profil des usagers. On observe, en effet, que statistiquement les contenus très techniques et peu vulgarisés sont peu denses en hyperliens (constitution européenne, RFC, …).

On peut aussi analyser les hyperliens contenus dans les documents en relation avec la future activité des usagers (Davison, 2002). On observe, par exemple, que 32 % des pages visitées par un usager médian correspondent à des liens contenus dans la page qu'il vient juste de visiter. Cette valeur passe à 47 % si l'on prend en compte les deux dernières pages visitées. En observant les choses autrement, on constate que les pages consultées par un usager médian ont 25 % de chances d'être consultées à nouveau. C'est-à-dire qu'en moyenne nous avons une chance sur quatre de revisiter

une page que nous avons déjà consultée. Si à présent, au lieu de considérer l'activité d'un usager nous observons celle d'un groupe, cette valeur passe à 50 %.

On pourrait s'interroger sur le sens de ce chiffre et sur ses facteurs d'évolution. Un groupe consulte, bien sûr, plus de pages qu'un usager isolé mais le taux de redondance de l'activité[1] étant rapporté au nombre total de pages consultées on aurait pu s'attendre à ce que les ratios restent proches quelque soit le nombre d'usagers considérés.

En réalité, le fait qu'un groupe engendre plus de redondances de consultation vient de l'effet de mutualisation et de synergie propre aux groupes. C'est-à-dire que le fait d'avoir plusieurs individus augmente la probabilité que l'un d'eux soit intéressé par les consultations des autres, augmentant par voie de conséquence le taux de redondance d'activité. La cohérence des groupes que nous avons évoqués au début de notre discussion est un facteur déterminant dans l'évolution de cette valeur. Cette observation nous a incité à proposer le taux de redondance de l'activité, facile à déterminer, comme une des mesures de cohérence des groupes. Un groupe composé d'individus ayant des centres d'intérêts très différents a de fortes chances d'avoir un niveau de redondance de l'activité commune plus faible qu'un groupe où les individus sont tous intéressés par le même sujet. Cette relation, relativement intuitive, a pu être vérifiée expérimentalement (Lancieri, 2005).

3.2. Structure d'occupation des espaces

Il est intéressant de faire le parallèle entre l'activité collective dans un environnement informationnel comme nous venons de le faire et dans un environnement physique. Pour approfondir cette question nous avons suivi les parcours d'une centaine d'usagers sur un site industriel (Benayoune et al, 2005). Les informations de déplacement ont étés obtenues grâce aux traces de passage entre 17 bornes WIFI reparties de manière homogène sur le site. Ces données sont disponibles en standard sur la plupart des routeurs WIFI (CISCO, etc.). La transformation logarithmique sur les deux axes de la relation liant le taux d'utilisation de chaque borne classée par ordre de fréquentation fournit une courbe dont la linéarité renvoie à la notion de similarité interne (Long tail, fractale,..) propre aux lois de puissances (e.g Pareto, Zipf, ...). Il est intéressant de noter que la pente de cette droite est un indicateur très synthétique de la structure d'occupation des espaces. Une pente proche de l'horizontale impliquerait une dispersion plus homogène sur toutes les zones de l'espace alors qu'une pente plus proche de la verticale serait le signe de la sur-occupation plus nette de certains lieux par rapport à d'autres (Lancieri, 2007).

[1] Le taux de redondance de l'activité de consultation est le ratio entre le nombre de pages uniques consultées rapporté au nombre total de pages consultées. Un taux de redondance élevé implique plus de réutilisation.

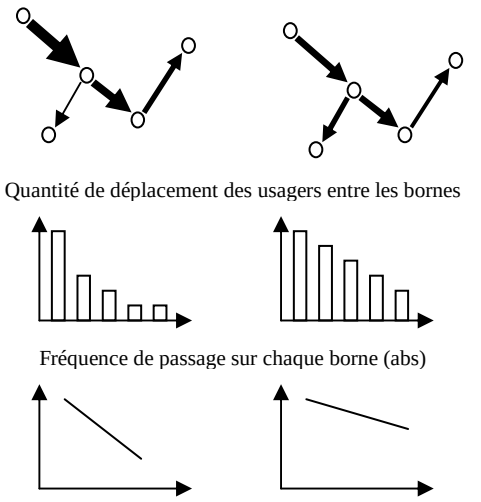

Figure 2 : Modélisation de la structure d'occupation des espaces.

Dans la figure qui précède, on observe ce principe mis en œuvre pour comparer des parcours surreprésentés dans certaines zones de l'espace (à gauche) par rapport à des répartitions plus équilibrés des chemins empruntés (à droite).

3.3. Modèle de comportement des usagers

Avec les exemples que nous avons présentés jusqu'à présent, les phénomènes étaient mesurés ponctuellement. L'évolution chronologique est difficile à évaluer avec ces méthodes qui exploitent essentiellement le calcul de similarités. Par exemple, pour identifier des groupes d'intérêts, il est possible de mesurer des distances entre profils d'usagers de manière à mettre dans les mêmes groupes des individus aux profils proches (clustering). Mais, pour mesurer l'évolution d'un profil dans le temps, il faut d'abord identifier sur quoi doit porter la mesure de similarité et par rapport à quoi il faut en apprécier l'évolution.

Pour clarifier le problème imaginons un profil utilisateur composé d'un vecteur de mots clés. Il est possible de représenter l'ensemble des mots clés des documents consultés sur une période plus ou moins longue, une sorte de moyenne. Imaginons maintenant que nous discrétisions l'information en capturant le profil construit sur une semaine toutes les semaines sur une période d'un an. Nous aurions 52 vecteurs qui nous permettraient d'avoir une vision de l'évolution des centres d'intérêt d'un individu au cours de l'année.

Le problème se situe dans le fait que cette information, quoique représentative, est très difficilement manipulable[2] et synthétisable car elle manque de référence. La méthode que nous proposons (Lancieri, 2005) permet de prendre en compte cette difficulté et simplifie grandement l'évaluation du comportement chronologique des individus. L'idée est de remplacer chacun des 52 vecteurs par une simple valeur numérique représentant à chaque semaine la distance entre le vecteur d'un individu et celui cumulant les vecteurs du groupe (la référence). Dans le temps cette distance va représenter l'évolution du comportement d'un individu comparé à celui du groupe.

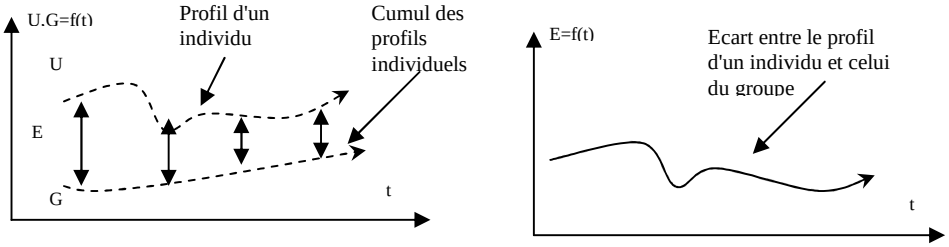

Figure 3: Evolution relative des comportements

Remarquons que l'inertie du profil de groupe est très importante car il cumule des comportements individuels en en lissant les variations. Comparée à l'inertie de groupe qui fait figure de point de référence par sa stabilité, le profil d'un usager va évoluer de manière plus importante. Par analogie, cette approche est semblable aux mesures de l'évolution des différences de potentiels en physique.

En plus de résoudre les problèmes que nous avons évoqués, cette approche a l'avantage de permettre l'utilisation directe des techniques du traitement du signal riches d'une longue expérience d'un point de vue méthodologique avec des bases théoriques éprouvées.

4. Réflexions sur les aspects éthiques et juridiques de l'utilisation de traces

Par définition l'utilisation de traces d'activité pose des problèmes de nature éthique et juridique. En effet chaque individu a un droit inaliénable à préserver sa vie privée et à ne pas dévoiler la part de son activité qu'il considère privée. Sur le plan juridique, la directive européenne "Data Protection Directive" et les législations des pays membres encadrent les pratiques (voir art L266.15 du code pénal). Le principe de base pour l'exploitation de données personnelles est finalement relativement logique dans le sens où l'usager doit être informé et consentant. Ceci étant dit, en dehors de l'union européenne les législations peuvent ne pas être cohérentes et la volonté de

[2] En effet chaque vecteur est un élément d'un espace multidimensionnel et peut être constitué de plusieurs dizaines voire centaines de mots clés pondérés par un facteur d'importance.

certains états à faire respecter la loi peut aussi être limitée quand des intérêts économiques et politiques sont en jeu.

De nombreux exemples ont montré que des traces pouvaient être utilisées à l'insu des usagers, avec parfois des conséquences douloureuses. Sans rentrer dans des débats philosophiques, nous pouvons faire plusieurs remarques sur ces points.

4.1. Des risques de natures diverses

Sur le plan technologique, et avant toutes autres considérations, il est important de faire le constat que la plupart des médias ont des capacités naturelles de traçage de l'activité. Initialement prévues pour le contrôle du bon fonctionnement des systèmes, ces capacités ont étés exploitées ensuite pour faire des statistiques d'usage puis pour prendre en compte le profil et le contexte d'activité des internautes (globalement ou de manière individuelle). Dans certains cas, la méconnaissance du média peut amener l'usager à certaines imprudences avec des conséquences inattendues. Aux Etats Unis, par exemple, certains internautes ont étés licenciés pour avoir mis sur leur blogs des critiques contre leurs entreprises ou leurs patrons. Pareillement, en France un proviseur a été révoqué pour avoir affiché son homosexualité sur son blog personnel (Nobs, 2006). En dehors des imprudences personnelles, le problème peut venir des imprudences de nos proches. En effet, certains services basés sur les réseaux sociaux (Upscoop, Facebook, Linked In, lastfm,..) exploitent la possibilité d'intégrer directement des carnets de contacts (eg. Gmail). Sauf à ne pas utiliser son e-mail ou à changer régulièrement d'adresse, ce qui revient au même (ne plus être joignable) chacun peut un jour découvrir son adresse sur des sites des plus improbables et se voir "arrosé" par des spams, simplement par l'imprudence d'un contact.

Les traces peuvent aussi être utilisées par des groupes de pressions pour orienter l'information, organiser le sur-classement de certains liens dans les moteurs de recherches ou faire des redirections abusives. Le cas du "google bombing" est assez typique de cette démarche qui a défrayé la chronique pendant la dernière campagne présidentielle en France. A cette occasion tout usager faisant une recherche sur "Nicolas Sarkozy" se voyait renvoyé vers un site en rapport avec "Iznogood". Exploitant une faiblesse de la technique du page-rank, la méthode employée est très simple et pratiquement imparable. Il a suffit qu'un nombre suffisant d'usagers mettent dans leur page web ou sur des blogs des liens entre le nom du candidat et l'adresse du site Iznogood-le-film. Comme on peut s'en douter, d'autres personnages politiques ont, avec plus ou moins d'humour, fait les frais de cette manœuvre (Georges Bush, …).

Ceci étant dit, les nuisances les plus importantes sont incontestablement liées à de la malveillance. Les spyware ou logiciels espions sont une des déclinaisons les plus courantes et des plus actuelles du risque associé à l'exploitation des traces. Ceci dit cette notion est floue dans l'esprit des individus et se mélange souvent à une forme

de psychose populaire qui déforme ou démultiplie la perception du risque. Un exemple de risque bien réel associé au spyware est par exemple le cas des keyloger qui permettent de capturer à distance ce que saisit l'usager (mot de passe, numéro de carte bleue, ...) pour en faire un usage détourné. Les keylogers peuvent prendre des formes[3] diverses, mais en général un firewall, un antivirus à jour et un peu de bon sens réduisent le risque de manière considérable. Par ailleurs, introduire ce type de processus sur la machine d'un utilisateur constitue un délit clairement identifié.

Ces exemples montrent que même si le risque associé à l'utilisation des traces existe et qu'il ne faut pas le sous estimer, il n'est pas toujours perçu dans sa réalité, tantôt surévalué et tantôt sous-estimé et pas forcément porté sur les bons acteurs.

4.2. Evolution de la perception du risque

Sur le plan historique, on peut constater que concernant l'activité médiatisée, la vision des frontières de la vie privée a beaucoup évolué au cours du temps. Au début de l'usage du téléphone, par exemple, les usagers considéraient comme une intrusion de devoir décrocher le combiné et de risquer d'avoir un interlocuteur avec qui ils ne souhaitaient pas dialoguer. Il était donc de coutume que les domestiques jouent le rôle de médiateur (Flichy, 2006). Plus récemment, le comportement des britanniques sur le plan des libertés individuelles a aussi évolué. Estimant d'abord que le simple fait de devoir présenter leur carte d'identité était une atteinte à leur liberté, ils acceptent aujourd'hui largement la vidéosurveillance qui envahit les villes anglaises (Porter, 2004).

Les raisons de ces changements d'opinions tiennent, sans doute, à la perception des avantages en termes de sécurité par rapport aux risques en termes de perte de liberté. La hausse de la criminalité, les attentats, etc, et l'usage on certainement modifié les données du problème en laissant apparaître que les anticipations négatives étaient peut être surévaluées. En ce qui concerne les traces d'activités réseaux, le problème peut aussi partiellement être abordé sous cet angle. Les arrestations régulières de terroristes ou de pédophiles grâce au suivi des traces sont aussi des éléments susceptibles de faire évoluer l'opinion publique. Dans un contexte proche, cette évolution de la perception de l'équilibre entre risques et avantages a pu être mesurée dans le domaine des transactions avec carte bancaire. La démocratisation de l'utilisation de la carte bancaire dans le commerce électronique ne s'est pas faite en un jour. Elle s'est installée après toute une époque de suspicions et de craintes. Avec l'habitude et la prise de conscience que le risque était limité par rapport aux avantages, les habitudes ont progressivement évolué.

Sous un autre angle, certains usagers ont vu un regain de la volonté hégémonique de Microsoft lorsque ce dernier proposa un contrôle des logiciels installés sur leurs machines. Avec le temps, même si ces craintes ne sont pas dissipées on s'aperçoit

[3] Voir aussi les screen scrappers, chevaux de Troie,..

que ce dispositif permet aussi de surveiller de manière relativement fiable le bon état de sécurité des machines en installant automatiquement des mises à jour qui sécurisent le système de manière importante. La question de l'équilibre entre les risques et les avantages se repose ici. En effet, même si on ne peut pas sous-estimer les arrière-pensées commerciales de Microsoft (ce qui peut paraître normal pour un industriel !), ce risque est il réellement plus important que celui d'avoir des trous de sécurité dans son système d'exploitation ?

Au-delà de la perception individuelle, la logique de groupe est aussi un facteur déterminant dans la mesure où de nombreux individus se déterminent par rapport à ce que fait la majorité. Ces effets moutonniers sont assez classiques et expliquent par exemple les excès des foules et autres bulles spéculatives dans les marchés financiers. La perception du risque dans les groupes est aussi impactée par ces phénomènes. Les éthologues pensent par exemple que l'organisation des animaux en groupes (banc de poisson, …) est partiellement motivée par le fait que la perception du risque d'être la victime diminue dans un groupe (Lenoir, 2004). Ce fait peut être constaté dans le cas du téléchargement illégal (peer to peer) ou l'argument du grand nombre apparaît plus ou moins consciemment comme une protection.

Quoi qu'il en soit, que l'on considère l'individu ou le groupe, les questions d'acceptabilité sont incontournables lorsque l'on cherche à mettre en œuvre un outil ou un service. Cette réalité se pose donc aussi en ce qui concerne l'exploitation des traces.

5. Exploitation des traces, des données aux services

Les traces d'activité peuvent être utilisées pour fournir des services à valeur ajoutée pour l'usager. En effet, les traces portent la connaissance dont l'extraction va alimenter le service. Cette connaissance va s'exprimer essentiellement par 2 notions apparemment sans rapport que sont le consensus et l'opposition.

Le consensus intervient en particulier dans les services de recommandation. Il s'agit de l'observation, faite par l'analyse de traces d'activités, que confrontée à un problème, à une question ou une situation particulière, une masse critique d'individus va adopter une position commune. Un exemple classique est celui d'un service qui, par exemple, conseille des livres de type B à un client qui recherche des livres de type A parce que le service a observé que la plupart des clients recherchant des livres de type A sont aussi intéressés par des livres de type B. L'évaluation et la notation des produits par les acheteurs rentrent aussi dans cette catégorie. L'acheteur en ligne averti exploite quasi systématiquement ces évaluation et conseils avant de faire son choix. On notera l'importance de la connaissance du contexte qui est nécessaire dans tous les cas. Le fait de savoir, par exemple, dans quelle tranche d'âge se situe le client pourra être une information utile pour recommander des CD lorsque la première requête est ambiguë et porte sur des contenus populaires qui revenant à la mode peuvent être appréciés par des publics de différentes générations. Le

contexte peut être fourni explicitement (déclaration par l'usager) ou fourni implicitement, et doit être reconstruit ou plutôt re-formalisé par recoupement des traces d'activité. Une combinaison de ces deux approches est aussi possible.

Dans le même esprit citons le cas d'un service permettant d'exploiter le consensus que nous avons proposé il y a quelques années (Lancieri, 1997, 2005). Ce système que nous avions baptisé miroir actif recyclait les contenus les plus souvent consultés par un groupe cohérent. L'idée de base était que dans ces groupes marqués par une proximité dans les objectifs, les intérêts ou les pratiques, les documents que certains trouvaient intéressants avaient de bonnes chances d'intéresser les autres membres du groupe. Les premiers résultats des expériences que nous avons menées sur ces thèmes se sont avérés très intéressants et ont confirmé l'importance de la contextualisation pour optimiser l'accès à la connaissance. Ces remarques restent vraies si on considère les aspects physiques, comme la recommandation de créneaux horaires dans un agenda ou de lieux pour établir une réunion.

Si l'exploitation de la position commune se révèle utile dans les environnements médiatisés, comme dans la vie courante d'ailleurs, on peut se demander comment la différence ou les oppositions peuvent être utiles. Pour répondre à cette question, il faut se rappeler que la connaissance naît de la différence pour paraphraser les propos de G. Bateson (Bateson, 1972). Ceci peut être illustré assez simplement en prenant l'exemple de l'accès à des contenus thématiques. Si on imagine que tout le savoir connu est réparti de manière thématique sur un certain nombre de sites web. Dans un premier cas, on observe qu'une population d'usagers réalise un nombre de consultations identiques sur chacun des serveurs. Quel que soit le nombre de consultations, fut-il infini, l'équi-répartition (le manque de différence) ne nous apprendra rien sur les centres d'intérêt du groupe. Cette connaissance commencera à émerger avec l'apparition de différences de consultations de certains thèmes par rapport à d'autres. Dans des travaux précédents (Lancieri, 2005), nous avons montré que ces écarts et les structures de connaissances pouvaient être modélisés de manière assez réaliste par des distributions sous exponentielles.

Citons un exemple concret pour montrer comment la connaissance peut venir de l'exploitation des différences. Googlefight est un service basé sur la recherche d'informations. Entre parenthèse, soulignons le fait que l'indexation des contenus du web jusqu'ici utilisé uniquement pour la recherche d'information devient la base de nombreux autres nouveaux services. Les opérateurs de moteurs de recherche ne s'y sont pas trompés et diffusent sous licence l'accès à leur contenu via une API, permettant ainsi à n'importe quel service d'exploiter leurs bases d'index. Googlefight est l'un de ces services. L'interface ressemble à celle d'un moteur de recherche classique, excepté le fait qu'il y deux champs de saisie au lieu d'un seul. Ainsi, en saisissant une expression par champ, ce service retourne la comparaison du nombre d'occurrences rencontrées pour chacune des deux expressions. A première vue cette application parait surtout ludique. En période d'élection on peut par exemple observer qu'un candidat est plus cité qu'un autre. On peut aussi avoir une idée de la

popularité de tel auteur, tel sport, telle idéologie par rapport à telle autre. En imaginant que le contenu du web est de plus en plus représentatif de ce qui intéresse l'humanité, les résultats de ce type de comparaison peuvent parfois être surprenants. Mais, au delà d'une utilisation ludique, ce service permet aussi de gagner du temps. Une utilisation assez classique est par exemple la vérification comparée de l'orthographe d'un mot, d'une expression, d'une traduction, etc. Citons aussi, dans un autre domaine, le cas des logiciels anti plagiat que certains enseignants commencent à utiliser tant le phénomène du copié-collé abusif depuis des sources trouvées sur Internet prends de l'ampleur.

5. Conclusion

Ces réflexions nous renvoient à notre discussion initiale sur le contexte et l'environnement. A l'époque où le temps et l'espace viennent étendre notre vision de la connaissance avec ses perspectives en termes de richesse mais aussi ses risques, l'exploitation des traces est au cœur de nombreux débats et de travaux de recherches.

En dehors des aspects méthodologiques, nous avons mis en avant deux points importants qui selon nous justifient d'investiguer ces thèmes. Le premier sur le plan de la modélisation permet d'envisager une meilleure compréhension des comportements individuels et collectifs. En se situant en complément des sciences sociales, l'analyse des grandes quantités de traces sur des longues périodes nous parait avoir un potentiel important. Le second point d'intérêt est l'exploitation des modèles issus de l'étude dans des services utiles aux usagers.

Ces avantages potentiels sont aussi accompagnés de risques qu'il ne faut pas négliger. Nous avons développé l'idée qu'une vision raisonnée de ces risques passait par l'information donnée aux usagers mais aussi par le débat public.

References

Porter, 2004 — Henry Porter; If you value your freedom, reject this sinister ID card; We should be afraid of future governments, whose nature we can't predict; The guardian; Friday December 17, 2004 http://www.guardian.co.uk/comment/story/0,,1375584,00.html

Davison, 2002 — Davison B.D Predicting Web Actions from HTML Content, Thirteenth ACM Conference on Hypertext and Hypermedia (HT'02)

Nobs, 2006 — Nouvel Observateur, Janvier 2006. Mende : le proviseur révoqué pour avoir affiché son homosexualité sur son blog; http://archquo.nouvelobs.com/cgi/articles?ad=societe/20060116.FAP0061.html&host=http://permanent.nouvelobs.com/

SVM, 2007 — Internet Menace-t-il notre vie privée? Dossier spécial Science et vie Micro octobre 2007.

Wiener, 1952 — Cybernétique et société (1952, rééd. 1971), Union Générale d'Editions, Collection 10/18

Doise, 2003	Willem Doise, Les relations entre groupes; dans Psychologie sociale, Livre collectif dirige par Serge Moscovici, 2003, Puf edit.
Derycke et al, 2007	Derycke Alain, Chevrin Vincent, Vantroys Thomas, P-learning in E-retail: a case study and flexible software architecture, Pervasive Learning 2007 workshop of the Pervasive 2007 conference
Ottavi et al, 2007	Alain Ottavi, Sylvain Baron, Luigi Lancieri, Capture et exploitation des traces dans un contexte de mobilité, Atelier Apprentissage Mobile, EIAH 2007
Benayoune et al, 2005	Toward a modelization of mobile learners behavior for the design and the evaluation of advanced training systems. Fares Benayoune, Luigi Lancieri, IADIS International Journal on WWW/Internet, Décembre 2005
Lancieri, 2001	The concept of informational ecology or interest of the information re-use in the company ; Luigi Lancieri ; ICEIS 2001; 3rd International Conference on Enterprise Information Systems; Setubal, Portugal;(IEEE, AAAi, TCNA)
Lancieri, 2005	Luigi Lancieri, Interactions humaines dans les réseaux ; Livre Hermes ed. ISBN 2-7462-1108-4 (May 2005)
Lancieri, 2007	Modelling collective behaviour using traces of individual activity; L..Lancieri; 19th International Conference on Systems Research, Informatics and Cybernetics (InterSymp2007)
Flichy, 2006	Patrice Flichy, Une histoire de la communication moderne, vie publique, vie privée, ed la découverte 2006, ISBN 2707126829
Bateson, 1972	Bateson, Gregory (1972). Steps to an Ecology of Mind: Collected Essays in Anthropology, Psychiatry, Evolution, and Epistemology. University Of Chicago Press. ISBN 0-226-03905-6. (en français, chez Seuil 1995)
Lenoir, 2004	Alain Lenoir, Evolution de la vie en groupe et de la socialite, Cours d'éthologie, 2004.
Pech, 2007	Marie Estelle Pech. Le copier-coller sur Internet irrite les profs, le figaro avril 2007, http://www.lefigaro.fr/france/20070410.FIG000000192_le_copier_coller_sur_internet_irrite_les_profs.html