Métrologie et modélisation de l'activité des forums de discussions

Luigi Lancieri
France Telecom R&D
Luigi.lancieri@francetelecom.com

Introduction

Les forums de discussions sont une forme d'applications de l'Internet dont les apports sont difficiles à évaluer de manière objective. Une des raisons souvent évoquée pour expliquer cette situation est le caractère informel et peu structuré du mode de contribution individuel qui semble un frein à une exploitation optimale de ce media. De fait, l'usage des forums parait limité à l'exploitation directe par des individus ou éventuellement par le biais d'outils de base comme les moteurs de recherche. Pourtant, il est clair que ce média, basé sur l'interaction entre individus, porteur d'intelligence collective, a un potentiel largement sous utilisé. En effet, il constitue un espace d'échanges riche couvrant de très nombreux sujets. Un autre point fort des forums réside dans la fraîcheur de l'information que l'on peut y trouver. Ce point dépend beaucoup de la popularité d'un groupe particulier mais constitue une différence par rapport à certains médias comme les sites Web qui sont des supports d'informations plus statiques.

En fait, le temps est un paramètre important pour apprécier les apports du modèle basé sur les forums (weblog, etc.) par rapport à celui basé sur les sites Web. On pourrait dire que la mise à jour des sites Web est d'autant moins fréquente que son contenu concerne des données stables. Par exemple un site sur les sciences fondamentales comme les théories mathématiques connues (i.e site d'éducation) ne nécessite pas de mises à jour fréquentes. A l'autre extrême un site de news (e.g l'actualité bibliographique concernant les mathématiques) contient souvent des données à la durée de vie courte et sera mis à jour avec une plus grande périodicité. Les forums échappent à cette logique car le taux de rafraîchissement du contenu est davantage lié à la popularité du forum qu'il implique ou non des données stables. Ceci est essentiellement dû au fait que le fonctionnement des forums reflète la dynamique des interactions entre individus.

La principale conséquence de cette forte dynamique est une transformation de la cinématique du mode de consultation qui passe du mode statique (i.e photo) au mode flux (i.e cinéma). Les posts anciens sont toujours accessibles (on parle de cinéma plutôt que de télé ou de radio) mais peu consultés car jugés peu actuels. Ce comportement se retrouve aussi dans d'autres medias, quand l'information est trop abondante, l'usager a tendance à trouver des biais cognitifs pour réduire l'espace de consultation (données plus récentes, recommandées, etc.). Pourtant la quantité n'est pas incompatible avec richesse, mais cette richesse ne deviendra vraiment exploitable que si l'on identifie des moyens d'automatiser le traitement de la connaissance.

Il existe 3 grandes catégories ou modèles d'usages de ce type d'espaces d'échanges. Le premier que l'on pourrait qualifier d'opportuniste implique un usage ponctuel. En cas de besoin, l'usager poste une question et consulte le forum pendant une période limitée avec l'objectif quasi exclusif de trouver une réponse à son problème. Le second usage implique un usage régulier mais en "spectateur" seulement. L'usager consulte régulièrement mais apporte peu de matière à la

communauté. Il navigue sur les forums comme il pourrait le faire sur le Web. La troisième catégorie implique un usage plus complet mais aussi plus localisé. L'usager est un habitué d'un nombre limité de forums mais il contribue en postant spontanément ou en répondant aux questions. Les différents types d'usages sont autant lies à la personnalités introverti ou non ou à la compétence de l'individu. Même s'il semble à première vue que le comportement des individus est finalement semblable à ce que l'on peut observer en société, il ne faut pas oublier l'impact de la technologie. En effet, l'anonymat relatif, par exemple, peut libérer des énergies et pousser un individu à s'exprimer alors qu'il ne l'aurait pas fait en public.

En résumant ces propos introductifs, on pourrait dire que divers éléments liés à la richesse des forums sont difficiles à exploiter car non disponibles formellement. En premier lieu, citons la connaissance comme contenu de la contribution individuelle ou collective des internautes mais aussi replacé dans un contexte épisodique. Ensuite, citons les données liées aux usages, aux comportements à l'intérêt des individus. Les modes d'interactions entre individus et leurs influences sur l'émergence de nouvelles connaissances sont aussi des éléments mal maîtrisés. Notre propos est que tous ces éléments de connaissances auraient intérêt à être capturés pour alimenter des grandeurs de caractérisation de manière à inférer des modèles cognitifs et sociaux. C'est l'objectif de la métrologie de l'activité des forums. Dans la suite de ce document, nous développons ce thème à la lumière de notre propre réflexion ainsi que par une synthèse d'autres travaux analogues.

Métrologie et modélisation

En offrant des outils de mesure des caractéristiques et de l'évolution des forums, la métrologie des groupes de discussions a 2 objectifs principaux: d'abord de permettre d'uniformiser et d'automatiser l'analyse et de traiter une grande quantité de forums aussi souvent que nécessaire. Ces informations permettent de mieux comprendre le mode de fonctionnement par nature complexe de ces environnements. En favorisant une "cartographie" des comportements et des interactions, ces données facilitent la prise en compte de modèles cognitifs et sociaux.

En plus d'indications sur la vie des forums ces informations peuvent être utilisées pour faire de la veille technologique dans une stratégie d'intelligence économique. Un certain nombre d'indicateurs quantitatifs d'activités peuvent être déterminés comme le niveau d'activité comparés par thèmes de discussions, ou le nombre moyen de posts par participants. Ce type de données peut être calculé de manière instantanée ou sur une période ainsi que sous l'angle de l'étude des variations. En terme d'application, ces données peuvent être utiles pour faire une sélection automatique des groupes actifs d'une manière générale mais aussi des pointes d'activités ponctuelles qui sont le signe d'un événement peu ordinaire ou de sujets intéressants faisant l'objet d'un débat particulièrement actif.

L'évaluation quantitative de l'activité (e.g. nombre de message par jour) est utile mais a ses limites. Par exemple, 10 messages par jours sur un groupe n'ont pas le même sens s'il s'agit de posts relatifs à un sujet nouveau (10 départs de file) ou s'il s'agit de réponses à un même message. Pour prendre en compte cet aspect plus qualitatif de l'activité du forum, il est possible de calculer un indice de nouveauté qui reflétera le niveau de jeunesse ou de maturité des discussions dans un forum. En effet, un forum qui ne produit que des nouveaux posts mais qui reste sans réponses a peu d'intérêt du point de vue de l'activité collective. A l'autre extrême, un forum qui ne produit que des réponses au même post finira par tourner en rond et produira peu d'innovation. Il existe sans doute un équilibre entre ces 2 extrêmes mais il est clair que son appréciation dépendra de l'individu qui selon le cas appréciera la profondeur des discussions nourries ou l'attrait de la

nouveauté. Le calcul du taux de posts nouveaux ne faisant pas référence à un post ancien (hors file) permet d'évaluer ce phénomène.

Cette grandeur a aussi l'intérêt d'évaluer la profondeur de la mémoire collective du forum. En effet, comme nous le disions, les usagers ont tendance à consulter les posts récents, considérant, parfois à tort, que les plus anciens sont périmés. En fait cette impression n'est pas toujours justifiée car l'expérience de la recherche d'information dans les posts anciens grâce aux moteurs de recherche, notamment, réserve quelquefois de bonnes surprises. Quoi qu'il en soit le fait de privilégier les posts récents a pour effet de jeter rapidement aux oubliettes ceux qui ne font pas l'objet de réponse. Une réponse en effet peut rendre actuel un poste parfois ancien. La mémoire du forum, c'est a dire l'ensemble des posts qui ont une bonne probabilité d'être consultés, (i.e. plutôt récents) sera donc fortement influencés par le niveau du taux de réponses.

Il est aussi possible d'envisager les groupes sous l'angle de l'interaction. Par exemple en déterminant le taux d'activité d'un usager habitué d'un groupe dans un autre groupe (cross-posting). Ce taux permet d'établir une cartographie des groupes ayant des relations basées sur l'usage effectif de leurs membres. Ces interactions réelles peuvent être rapportées à la proximité thématique de chaque groupe. Ces 2 paramètres peuvent être associes au sein d'un indicateur synergie potentielles. Dans ce contexte, nous avons proposé [11] une méthode pour évaluer le niveau d'interactions dans les groupes de discussions. La métrique associée basée sur la puissance itérée (formulation issue de la théorie des graphes) permet avantageusement de prendre en compte les contributions indirectes dans les faisceaux d'interactions. La valeur globale obtenue synthétise les diverses contributions dans les groupes et en particulier dans des situations de cross posting. Cette méthode peut aussi être employée à l'intérieur des groupes de grandes dimensions où une segmentation artificielle a été opérée (algorithmes de catégorisation). La contribution croisée des usagés aux différents segments peut aussi être capturée et formalisée par cette méthode. Nous avons aussi proposé [12] une méthode de mesure du niveau de coopération potentiel entre individus basée sur le niveau de recouvrement de leur "profil thématique" (mots les plus fréquents rencontrés dans les posts d'un usager sur une période). Ces grandeurs sont intéressantes pour suivre l'activité des forums: évolution dans le temps du niveau global de synergie potentielle mais aussi pour être associées à des modes de mise en relation. Par ailleurs, les modèles statistiques liés à certaines grandeurs caractéristiques des forums mettent en évidence une propriété self similaire. Le modèle de loi de probabilité sous exponentiel associé (Pareto, Weilbull, log normal, etc) dit à mémoire par opposition à des modèles de type loi de Poisson qui ignore "le passé" a dans de nombreux travaux été mis en avant pour refléter le comportement humain.

Etat de l'art

Nous reportons dans cette section un certain nombre de travaux liés à la métrologie des forums. Saito [9] présente une étude statistique des groupes de news sous l'angle technologique. Il présente d'abord les protocoles et les architectures de service associés aux serveurs de news (nature des échanges, organisation des réseaux, format et contenu des traces, etc). En dehors de résultats quantitatifs (nombre de posts par groupes, distribution des tailles d'articles, etc), des conclusions de type comportementales sont inférées par exemple par la comparaison des groupes en fonction du nombre de message ou de la taille des messages, montrant que certains groupes évoluent vers des échanges plus multimédia.

Le projet NetScan [13] financé par Microsoft permet d'obtenir de nombreux détails quantitatif sur le comportement des groupes Usenet. Les chiffres mis a jour mensuellement sur tous les groupes (recherche des groupes par mots clés). On y trouve des grandeurs comme le nombre de messages

sans réponses, la taille des messages, le nombre de personnes ayant répondu à au moins un messages, etc. Au total 9 grandeurs permettent de faire des comparaisons et de quantifier les comportements des groupes.

Une des difficultés freinant l'étude fine du comportement des usagers est l'obtention de traces fines de leurs activités. Pour palier cette contrainte, certains auteurs ont modifié les outils de consultation (news readers) de manière à obtenir des traces plus détaillées. D. Maltz [1], par exemple, a modifié les news readers NN et XRN de manière à obtenir des traces sur les sessions de consultation. Des données comme la durée de la session (ouverture, fermeture du news reader), le temps passé a scanner les sujets ainsi que le temps passé a lire un article (temps passé entre ouverture et fermeture de l'article) deviennent accessibles. L'étude met en évidence plusieurs résultats intéressants sur le comportement des usagers. L'auteur étudie par exemple le ratio entre le nombre de groupes consultés par session et le nombre de groupes auxquels se sont inscrit les usagers montrant que ce ratio décroît très vite (log normal) avec le nombre d'inscriptions ce qui implique que les usagers qui s'abonnent à de nombreux groupes ont tendance à surestimer leurs capacités de lecture. La moitié des usagers s'abonne à moins de 20 groupes et arrive dans 90 % des cas à les suivre. La fraction des usagers abonnés à plus de 100 groupes est de 8 % alors que seul 1 % arrive à suivre.

Dans le même esprit, Jones et al [2] ont étudié via une approche statistique les phénomènes de communications de masses dans les communautés électroniques (études de 578 groupes pendant 5 mois). Les auteurs ont cherché à étudier l'impact des limites du processus cognitif dans les communications de groupes en mesurant les interactions entre utilisateurs. Ils commencent par proposer une méthodologie pour reconstituer automatiquement les files d'échanges sur Usenet, ce qui n'est pas simple compte tenu de la diversité des outils de consultation. En travaillant sur les indicateurs internes aux messages (Re, Reply, <, etc) ils construisent un modèle basé sur les probabilités conditionnelles permettant d'évaluer automatiquement qu'un message est effectivement une réponse. Cette méthode permet de reconnaître automatiquement une réponse dans 99 % des cas. Ces études ont permis d'évaluer un certain nombre d'hypothèses. Elles ont montré que la complexité des messages (nombre de lignes) avait tendance à réduire fortement (loi sous exponentielle) dans les groupes très actifs (nombreuses personnes postant des messages). Les mesures de régression confirment la fiabilité du modèle (le nombre de posteur permet de prédire le nombre de ligne des posts). Les auteurs ont aussi découvert que les messages qui démarraient un fil de discussion avaient tendance à être courts (peu complexe). Le pouvoir prédictif de ce modèle a été démontré dans 63 % des cas. Un autre point interessant est que la participation des usagers est moins stable (usagers moins impliqués) dans les groupes très actifs où l'activité s'explique plus par une faible participation de nombreux intervenants que par des participations individuelles soutenues. Les travaux qui ont suivi [3] cette étude, ont aussi montré que le type de technologie a un impact sur cette stabilité des contributions. L'auteur est arrivé à cette conclusion en comparant la stabilité des contributions dans le mode "groupes de news" et dans le mode "listserv". Dans ce dernier mode les contributions sont plus stables (50 % des usagers postent des messages sur une période de 2 mois (i.e les usagers ayant posté un mois ont posté le mois suivant) contre 11 % pour les groupes de news.

Wittacker et al [5], quant à eux, ont étudié l'évolution de 500 groupes de discussion pendant 6 mois. Ils ont réalisé des statistiques de manière à mettre en évidence différentes caractéristiques des groupes sous l'angle des interactions de masses. L'étude a débuté par une évaluation démographique de la population concernée (nombre de posteurs, nombre de messages par posteur, intervalle entre posts). Les auteurs font ressortir un indicateur décrivant l'évaluation du niveau de familiarité d'un posteur avec un groupe (27 % des posts proviennent de personnes n'ayant contribué qu'une fois alors que seul 2.9 % des usagers engendrent 25 % des posts). Le nombre

moyen de contributions par utilisateur est de 3.1. Les auteurs n'y font pas allusion mais la courbe nombre de messages /posteur a l'allure classique d'un fonction sous exponentielle. En terme de structure d'interaction (stratégie d'échanges), l'étude révèle que 34 % des messages de chaque groupe sont aussi adressés à au moins un autre groupe (3.1 groupes en moyenne). On observe aussi que sur le plan du cross posting chaque groupe a été en relation avec 272 groupes différents avec une cohérence faible (5.4 post par groupe en moyenne, i.e nombre de posts depuis l'usager habituel d'un groupe vers un groupe donné). Les auteurs ont aussi mesuré le niveau d'interactivité en étudiant la profondeur des files. Ils montrent qu'en moyenne une file contient 1.8 messages alors que 33 % des files contiennent plus de 2 messages (extension de conversation réussie). Les initialisations de communications manquées (messages seuls dans leur file) correspondent à 44 % des messages. Les auteurs ont mesuré que 54 % des groupes produisaient des FAQs (Frequently Asked Questions) vus comme productions de connaissance plus structurées par le groupe. Les auteurs ont ensuite étudié l'impact de différentes variables sur l'évolution du groupe. La croissance de la production de FAQs, la décroissance du cross posting, la croissance de la taille des messages sont vues comme des facteurs d'augmentation de dénominateurs communs ou de la cohésion du groupe (common ground). L'auteur a utilisé un modèle causal construit à partir d'analyse de régression entre les variables. Il déduit par exemple que les groupes de grande taille ont tendance à engendrer du cross posting et à contenir des messages de faible taille. Pareillement les groupes contenant beaucoup d'usagers familiers ont tendance à avoir moins de cross posting et contenir des messages plus longs. Par contre la familiarité ne semble pas avoir d'effet sur la production de FAQs qui est surtout liée à la présence d'un modérateur. Les auteurs tirent diverses conclusions. D'abord, que même si tout le monde peut poster dans un news group, les messages sont surtout générés par une minorité très active (bavarde). Ceci contraste avec la communication plus physique (face à face ou via la vidéo ou l'audio seul) qui implique, toute proportion gardée, une forme de communication plus égalitaire.

Smith [7] a étudié les structures sociales émergentes ou invisibles des groupes Usenet. Au travers d'un rappel sur les racines et l'état actuel de Usenet (histoire, population, popularité par type de groupes, répartition géographique, organisation, technologie). L'auteur présente ensuite des statistiques détaillées sur les caractéristiques des messages, des usagers ou de la répartition chronologique des posts.

Viagas et al [6] ont étudié la contribution individuelle des auteurs à l'effort collectif ainsi que le positionnement de cette participation individuelle dans le temps. Différentes métriques et des représentations graphiques associées ont permis aux auteurs d'inférer un niveau de confiance à chaque contributeur en fonction de son niveau d'activité ou par croisement avec d'autres contextes où il intervient. Des études d'usages ont aussi montré que ces indicateurs étaient utilisés par les usagers pour choisir les messages à lire (modification des usages). Les auteurs ont obtenu les données issues de l'activité des groupes Usenet dans le contexte du projet NetScan [13][7] L'analyse visuelle des formes graphiques issues des remontées statistiques et chronologiques de l'activité a permis d'inférer des comportements de type initiateur de discussions, adepte des débats, contributeur de fonds ou perturbateur spammeur. Une approche analogue a été approfondie dans un contexte de navigation Web [8].

D'autres travaux n'impliquant pas seulement les forums mais aussi d'autres modes de communication ont montré l'impact de la modalité sur la coopération. Jensen et al [4] ont montré que la contribution des usagers à une activité en commun (un jeu en ligne) était plus importante quand le mode de communication était plus évolué (voix, synthèse vocale, texte chat, pas de communications). Les résultats montrent que le niveau de participation est croissant avec le niveau du mode. Les auteurs attribuent cette relation au niveau de confiance qui s'instaure avec

des formes plus évoluées de communication. On notera que la synthèse vocale engendre plus de contributions que le texte chat équivalant mais moins que la conversation humaine.

Références

[1] David A. Maltz. Distributing Information for Collaborative Filtering on Usenet Net News. SM Thesis, Massachusetts Institute of Technology, Cambridge, MA. 1994. Available as MIT/LCS/TR-603 and Xerox PARC CSL-94-5.; http://www.lcs.mit.edu/publications/pubs/pdf/MIT-LCS-TR-603.pdf

[2] Jones, Q., Ravid G., and Rafaeli S. (2002). " An Empirical Exploration of Mass Interaction System Dynamics: Individual Information Overload and Usenet Discourse." In: Proceedings of the 35th Annual Hawaii International Conference on System Sciences, IEEE, Big Island, Hawaii. http://modiin.njit.edu/docs/hicss%20on%20system%20dynamics.PDF

[3] Jones, Q., (2003) "Appling Cyber-Archaeology", Proceedings of the Eighth European Conference on Computer Supported Cooperative Work, 14–18 September 2003, Helsinki, Finland. Kluwer Academic Publishers, Dordrecht Hardbound, ISBN 1-4020-1573-9. http://modiin.njit.edu/docs/jones_ecscw03.pdf

[4] Jensen, C., Farnham, S., Drucker, S., & Kollock, P. The Effect of Communication Modality on Cooperation in Online Environments. In Proceedings of CHI 2000, The Hague, Netherlands March 2000. http://research.microsoft.com/scg/papers/dilemmachi2000.pdf

[5] Whittaker, S. Terveen, L., Hill, W., and Cherny, L. (1998). The dynamics of mass interaction, In Proceedings of Conference on Computer Supported Cooperative Work, 257-264. New York: ACM Press. http://dis.shef.ac.uk/stevewhittaker/cscw98-published.pdf

[6] Fernanda B. Viégas; Marc Smith; Newsgroup Crowds and AuthorLines: Visualizing the Activity of Individuals in Conversational Cyberspaces; Proceedings of the 37th Hawaii International Conference on System Sciences – 2004; http://research.microsoft.com/~masmith/Newsgroup%20Crowds%20and%20Author%20Lines.pdf

[7] Smith, Marc. "Invisible Crowds in Cyberspace: Measuring and Mapping the Social Structure of USENET" in Communities in Cyberspace, edited by Marc Smith and Peter Kollock. London, Routledge Press, 1999; http://research.microsoft.com/~masmith/Invisible%20Crowds%20in%20Cyberspace.doc

[8] Anne Lavallard et Luigi Lancieri; Observation de l'évolution des communautés d'intérêts, IC 2004 (15 em journées francophone d'ingénierie des connaissances, Lyon mai 2004.; http://www.ensicaen.ismra.fr/~lancieri/public_html_fichiers/IC2004.pdf

[9] Yasushi Saito, Jeff Mogul, and Ben Verghese; A Usenet Performance Study"; Sep 1998. project report; HP Labs. http://www.hpl.hp.com/personal/Yasushi_Saito/usenet.ps

[10] Gary Flake, Steve Lawrence, C. Lee Giles, Efficient Identification of Web Communities,(KDD 2000) http://www.neci.nec.com/~lawrence/papers/web-kdd00/web-kdd00.pdf

[11] Luigi Lancieri; A connectionist approach for evaluating the complexity of interactions in the World Wide Web. The case of News Groups ; IEEE International Joint Conference on Neural Network 2000 IJCNN 2000- (COMO - Italy); http://www.ensicaen.ismra.fr/~lancieri/public_html_fichiers/ijcnn00.pdf

[12] Luigi Lancieri; Reusing Implicit Cooperation, A novel approach to knowledge management; In tripleC (Cognition, Cooperation, Communication) International Journal, 2004 pp 28-46; http://www.ensicaen.ismra.fr/~lancieri/public_html_fichiers/tripleC2(1)_Lancieri.pdf

[13] Usenet social accounting search engine; Microsoft; http://netscan.research.microsoft.com/