Outil de représentation des évolutions de communautés d'intérêts

Anne Lavallard, Luigi Lancieri

France Telecom R&D
42 Rue des coutures 14000 Caen
anne.lavallard@francetelecom.com
luigi.lancieri@francetelecom.com
http://www.ensicaen.ismra.fr/~lancieri

Résumé. Cet article présente un système de visualisation permettant l'observation des comportements collectifs implicites. Il s'agit de reconnaître et de représenter des communautés à partir des connexions Internet des utilisateurs : les utilisateurs sont répartis en communautés en fonction des similarités entre des listes de termes établies sur l'analyse des documents consultés par chacun d'eux. L'étude est rendue dynamique par la comparaison des communautés reconnues sur des périodes de temps connexes. L'outil décrit ci après offre deux représentations différentes de ces communautés : une vision des liaisons thématiques entre les utilisateurs sur chaque période étudiée et une vue comparative des communautés reconnues sur toute la durée de l'étude.

1 Introduction

La problématique de cette étude est de proposer des outils de visualisation de la dynamique des communautés d'intérêts. Même si des algorithmes permettent de calculer des communautés thématiques en analysant les documents consultés par les utilisateurs, la lisibilité des résultats est parfois faible. En effet, d'une part, il est difficile de prendre en compte certains paramètres comme l'aspect dynamique du comportement. D'autre part, les résultats sont parfois difficilement exploitables directement, faute de représentation visuelle adaptée pour guider l'interprétation des données.

Pour répondre à cette problématique, nous avons mis au point différents outils afin de pouvoir identifier et visualiser différents paramètres liés à l'évolution des groupes d'usagers. L'accent a été porté sur l'aspect temporel des données observées. Le but final étant de permette une observation optimisée des comportements collectifs par la recherche des propriétés dynamiques liant des groupes d'usagers (communautés implicites). L'identification et la visualisation de ces propriétés sur des périodes de temps connexes permet par confrontation d'évaluer les évolutions dans le temps de ces communautés.

Cette étude a été réalisée sur la base de données d'activités réelles des usagers sur une période de 17 mois (traces de consultations et analyse des documents associés). Ces données ont étés rendus anonymes pour concilier notre souci de préservation de la vie privée avec notre approche non intrusive de l'observation de l'activité.

Outre l'apport du point de vue de l'ergonomie de la représentation de l'activité collective, ces premiers travaux ont permis d'ouvrir des nouvelles pistes de réflexions permettant d'affiner l'analyse des résultats. Cet objectif est particulièrement important dans un contexte d'optimisation de l'exploitation des phénomènes d'intelligence collective.

2 Etat de l'art

2.1 Visualisation de données

La visualisation doit pouvoir permettre à un acteur humain de comprendre le système qui lui est présenté. Cette présentation doit faciliter la tâche de l'observateur afin qu'il puisse mobiliser toutes ses ressources à l'interprétation des données. C'est pour cela que presque tous les systèmes de visualisation offrent des représentations spatiales [Bertin, 1977]. Il faut également veiller à ne pas perdre l'observateur dans une profusion de détails : une solution est donc de présenter en premier lieu une vision globale peu détaillée, tout en lui permettant de choisir une zone qui sera affinée selon ses souhaits [Scheiderman, 1996]. De plus, en détaillant la vue, il peut être très appréciable de conserver une vision, même déformée, du contexte. Enfin, dans toutes les manipulations et les déplacements de l'observateur dans le système, il faut veiller à favoriser l'interactivité sur les données ainsi que la continuité entre chaque vue, toujours dans le but de focaliser l'attention sur l'interprétation sans qu'elle ne butte sur des points de représentation.

2.2 Différents systèmes existants

Les représentations axiales [Mackinlay et al., 1991] permettent d'ordonner les données suivant un axe privilégié, en général, le temps. Il est donc possible d'observer d'une part les évolutions de chaque élément, et d'autre part, de comparer les éléments placés en parallèles.

Pour la présentation d'arbres, les techniques peuvent être séparées selon leurs approches :

- Approche diagramme [Lamping et al., 1995][Battista et al., 1999][Herman et al., 2000]
- Approche surfacique [Johnson et Schneiderman, 1991][Andrews et Heidegger, 1998]
- Approche en trois dimensions [Robertson et al., 1991][Munzner, 1997]

Les structures en graphes, ou réseaux, sont également très courantes. Elles sont cependant plus difficiles à représenter avec clarté. La plupart des algorithmes utilisent une métaphore mécanique pour calculer la place de chaque nœud pour une lisibilité optimale [Kamada et Kawai, 1989].

Des techniques comme la vue en œil de poisson [Furnas, 1986] permettent de visualiser les détails d'une zone particulière tout en conservant une vue -déformée- de l'ensemble.

2.3 Bilan

Au cours de l'élaboration de cet état de l'art, nous avons pu mesurer l'importance du choix de la structure à mettre en valeur sur la représentation. En effet, les données sont souvent polymorphes, et peuvent être visualisées de différentes façons en fonction des éléments que l'on souhaite faire apparaître.

Dans le cadre de nos travaux, nous cherchons à visualiser des communautés, ainsi que leurs évolutions. Deux structures principales ont été choisies : un aspect réseau entre les utilisateurs, qui nécessite une représentation sous forme de graphe et un aspect linéaire, le long de l'axe du temps, pour pouvoir y comparer les caractéristiques des communautés reconnues.

Deux modes de visualisations ont donc été construits autour de ses deux approches, et introduits dans le calcul à deux stades différents.

3 Méthode et description

3.1 Etude des données

Les données obtenues par nos travaux de recherche sur l'évolution des communautés sont trop spécifiques pour y adapter des systèmes de visualisation standard.

La recherche de communautés se fait sur des traces consultations de sites Web. Pour introduire une dynamique dans cette étude, ces traces ont été divisées en différentes périodes selon la date de connexion. Par ailleurs, pour chaque document consulté, on en extrait les termes les plus fréquents qui seront par la suite considérés comme la signature thématique du document.

Pour chaque période d'étude et pour chaque utilisateur, il est donc établi une liste de termes qui caractérise son comportement durant cette période. Cette liste permet d'introduire une métrique entre les utilisateurs, de les situer les uns par rapports aux autres, puis de les regrouper en communautés [Lancieri et Durand, 2003]. L'étude a été réalisée sur une période de 17 mois d'activité de 320 utilisateurs.

Nous avons donc à cette étape, c'est-à-dire avant le calcul effectif des communautés, les distances entre chaque utilisateur que nous pouvons réunir dans une matrice de distance. Il y a une matrice de distance par période étudiée.

3.2 Le Graphe des utilisateurs

Pour cette représentation, l'espace central (A) est réservé à la présentation d'une période particulière (la *FIG. 1* correspond à la période 4). Dans cet exemple, chaque période correspond à environ deux mois d'activité. L'aspect dynamique est donné par une série de vignette en bas qui donnent une schématisation des autres périodes, par des outils de navigation, ainsi que par des repères colorés.

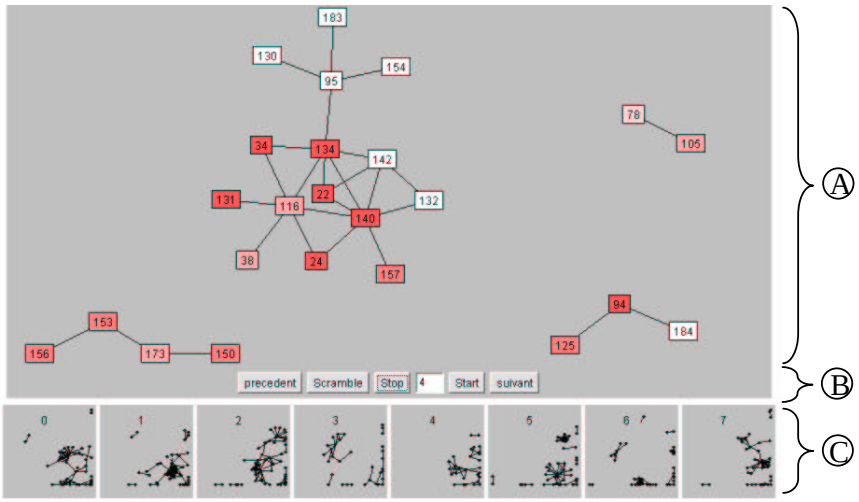

FIG. 1 - *Graphe des utilisateurs.*

3.2.1 Représentation d'une période

Sur le graphe des utilisateurs (A), chaque sommet représente un utilisateur, et une arête représente une liaison entre deux utilisateurs. On estime que deux utilisateurs sont liés si la distance qui les sépare est inférieure à un certain seuil paramétrable. Les utilisateurs non reliés ne sont pas représentés. La longueur des arêtes est dans la mesure du possible proportionnelle à la distance correspondante.

Le calcul de leur position se fait à l'aide de la métaphore mécanique : les sommets placés aléatoirement au départ vont se rapprocher progressivement des sommets desquels ils sont liés, jusqu'à obtenir un optimum. Le système de forces peut être arrêté ("stop") et relancé ("start") par l'observateur. Il peut également déplacer les sommets à la souris.

3.2.2 Représentation du contexte dynamique

Le contexte dynamique est maintenu par une série de vignettes (C) qui présente l'état des autres périodes. La navigation dans le temps s'effectue à l'aide des boutons précédent et suivant, ou bien en indiquant directement le numéro de la période souhaitée (B). Une gestion en parallèle de tous les affichages de toutes les périodes permet de conserver une continuité dans cette navigation.

La reconnaissance des communautés, c'est-à-dire des groupes d'utilisateurs qui perdurent dans le temps, est facilitée par le marquage rouge : en effet, un individu est indiqué en rouge s'il était présent dans une ou plusieurs périodes précédentes, et l'intensité du rouge augmente avec le nombre de périodes successives précédentes sur lesquelles il apparaît. Un groupe d'utilisateurs présent en rouge foncé correspond donc à une communauté bien établie.

3.2.3 Conclusion

Ce système permet donc d'avoir une représentation des données avant tout nouveau calcul. Les communautés ne sont pas déterminées explicitement, mais l'observateur peut en apprécier les évolutions. Dans nos travaux, nous avons remarqué deux types de communautés qui semblent intéressants : des groupes présentant un sous graphe presque complet, ou bien des groupes centrés autour d'un individu.

3.3 Calcul des communautés

Un algorithme de clustering agglomératif [Ronkainen, 1998] détermine, à partir de la matrice des distances, pour chaque période, la répartition des utilisateurs dans les différents groupes. Ensuite, pour reconnaître les communautés, on cherche les groupes d'utilisateurs relativement stables d'une période à une autre. C'est le pistage des communautés. On notera l'intérêt d'une étude comparée des ces communautés implicites stables avec les communautés déclarées ou imposés (groupe de travail, équipe de recherche, etc.).

A l'issue de ce calcul, nous disposons donc pour chaque communauté reconnue de sa composition en membres à chaque période. Il est également possible d'obtenir les termes communs à tous ses membres.

3.4 La Frise des communautés

Nous souhaitons ici représenter le comportement d'une communauté sur l'axe du temps. Elle est figurée par l'histogramme de l'évolution du nombre de ses membres (*FIG. 2*). Des courbes pourraient être ajoutées pour visualiser d'autres caractéristiques, comme le nombre de membres permanents ou le taux de renouvellement. Les membres principaux et les termes principaux de la communauté pointée par la souris sont indiqués à droite.

Une étude sur l'association de couleur en fonctions des termes émergents est en cours. Elle permettrait de visualiser une évolution thématique d'une communauté au travers d'un changement de couleur.

Les différentes communautés reconnues sont ainsi placées les unes en dessous des autres. Elles sont classées en fonction de leur importance. Cette importance est calculée à partir de sa durée de vie et de son nombre moyen de membres. Le réglage de la hauteur permet d'afficher plus ou moins de communautés.

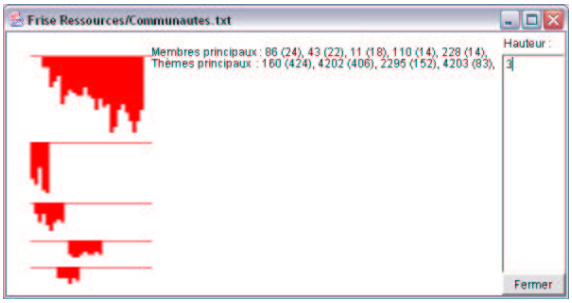

FIG. 2 - *Frise des communautés.*

Ce système permet une représentation synthétique des communautés. Il nous a permis d'avoir une vision plus précise de la dynamique communautaire. Par exemple, des communautés distinctes ont présenté les mêmes membres principaux et les même thèmes principaux : on peut donc supposer que les deux groupes ainsi présentés n'étaient que deux facettes d'une même communauté.

4 Conclusion

Dans le cadre de notre étude sur l'évolution des communautés d'intérêts sur des espaces d'activités réels continus, les systèmes de représentation visuels sont des outils importants. Ils ont permis de concrétiser des résultats jusqu'à présent abstraits, de percevoir des particularités invisibles par le biais des statistiques. Ils permettent également une certaine validation des calculs et un ajustement des algorithmes. A terme, ils doivent nous permettre de dégager des propriétés générales des communautés : évolutions en membre, évolution thématique, durée de vie, etc.

Les deux méthodes de visualisations décrites sont intégrées à une interface qui permet de fixer certains paramètres avant d'effectuer les calculs. Puis elle permet d'afficher la représentation des résultats correspondants. Il est ainsi possible d'observer l'impact de ces paramètres sur les résultats.

Références

[Andrews et Heidegger, 1998] K. Andrews et H. Heidegger. Information Slice: Visualising and Exploring Large Hierarchies using Cascading, Semi-Circular Discs. *IEEE Symposium on Information Visualization (infoVis'98),* 1998.

[Battista et al., 1999] G.D. Battista, P. Eades, R. Tamassia et I.G. Tollis. *Graph Drawing: algorithms for the visualization of graphs*. Prentice Hall, 1999.

[Bertin, 1977] J. Bertin. *La Graphique et le traitement Graphique de l'information*. Flammarion, 1977.

[Furnas, 1986] G. Furnas. Generalized Fisheye View. *ACM Conference on Human Factors in Computing Systems, CHI'86 Conference,* Boston, MA, 1986.

[Herman et al., 2000] I. Herman et al. Graph visualization and navigation in information visualization. *IEEE on Visualization and computer Graphics,* 2000.

[Johnson et Schneiderman, 1991] B. Johnson et B. Schneiderman. Tree-maps: A space-filling approach to the visualization of hierarchical information structures. *IEEE Visualization ' 91,* San Diego, CA, 1991.

[Kamada et Kawai, 1989] T. Kamada et S. Kawai. An Algorithm for Drawing general Indirect Graphs. *Information Processing Letters,* 1989.

[Lamping et al., 1995] J. Lamping, R. Rao et P. Pirolli. A Focus + Context Technique Based on Hyperbolic Geometry for Visualizing Large Hierarchies. *ACM Conference on Human Factors in Computing Systems, CHI'95 Conference,* Denver, CO, 1995.

[Lancieri et Durand, 2003] L. Lancieri et N. Durand. Evaluating the impact of the user profile dimension on its characterization effectiveness. *IEEE Symposium on Computational Intelligence for Measurement systems and applications, (CIMSA2003),* 2003.

[Mackinlay et al., 1991] J.D. Mackinlay, G.G. Robertson et S.K. Card. The Perspective Wall: Detail and Context Smoothly Integrated. *ACM Conference on Human Factors in Computing Systems, CHI'91 Conference,* New Orleans, LA, 1991.

[Munzner, 1997] T. Munzner. H3: Laying Out Large Directed Graphs in 3D Hyperbolic Space, *IEEE Symposium on Information Visualization (infoVis'97),* 1997.

[Robertson et al., 1991] G.G. Robertson, J.D. Mackinlay et S.K. Card S. Cone Trees: Animated 3D Visualizations of Hierarchical Information, *ACM Conference on Human Factors in Computing Systems CHI'91 Conference,* New Orleans, LA, 1991.

[Ronkainen, 1998] P. Ronkainen. *Attribute Similarity and Event Similarity Sequence in Dated Mining,* Technical C-1998-42 Carryforward, University of Helsinki, 1998.

[Shneiderman, 1996] B. Shneiderman. The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations. *IEEE Visual Languages,* 1996.

Summary

This paper presents a visualization tool developed for the study of the implicit collective behaviors of Internet users'. Communities are identified and visualized from the Internet users' accesses and consulted textual documents. The users are distributed in the communities according to the similarities between the set of most frequents words extracted from their consulted web pages. Furthermore, the comparison between the communities trough different period gives an overview of the collective dynamic behaviors. This tool offer two way of representation: the first shows the thematic links between the users and the second shows a global view of the communities' evolutions.