Observation de l'évolution des communautés d'intérêts

Anne Lavallard * et Luigi Lancieri **

France Telecom R&D, 42 Rue des coutures, 14 000 Caen
* anne.lavallard@francetelecom.com
** http://www.ensicaen.ismra.fr/~lancieri, luigi.lancieri@francetelecom.com

Résumé

Cet article présente un outil permettant de visualiser les évolutions de communautés d'intérêts. Ces communautés sont calculées à partir des connexions Internet des utilisateurs : l'indexation des documents consultés permet le regroupement des usagers en communautés. L'accent est porté sur les transformations que subissent ces groupes au cours du temps.

Mots clés : Ingénierie des connaissances ; visualisation, communauté, centre d'intérêts

1 Introduction

Sur un réseau local, il est fréquent que plusieurs personnes cherchent sur Internet des informations thématiquement proches. Nous qualifions ces groupes de communautés d'intérêts implicites dans la mesure où ils sont stables dans le temps.

En dehors de l'intérêt que représente une meilleure perception des interactions informationnelles dans l'entreprise, la connaissance de l'existence de ces communautés implicites ouvre la porte à de nombreux services. Citons par exemple, la suggestion ciblée d'information.

Le but de notre outil est d'analyser les connexions Internet des utilisateurs, périodes par périodes, de les regrouper suivant leur proximité thématique et d'observer les évolutions des communautés ainsi formées par une représentation de leurs principales caractéristiques dynamiques.

2 Calcul des communautés

2.1 Données utilisées

La maquette fonctionne aujourd'hui sur les traces d'activités réelles de l'ensemble des utilisateurs d'un intranet regroupant 300 utilisateurs sur une période de 17 mois.

Cette période d'étude a été sectionnée en 32 segments de 2 mois se chevauchant par trois-quarts. Le chevauchement des périodes permet d'accentuer la redondance des informations et ainsi de mieux suivre les évolutions des relations thématiques entre les utilisateurs.

Un premier travail [1] nous a permis de visualiser, période par période, les relations thématiques entre utilisateurs (graphes). Mais pour mieux suivre les évolutions internes d'une communauté et les comparer entre elles, (composition, stabilité des membres, etc) il était nécessaire de trouver une autre forme de représentation.

2.2 Distances entre utilisateurs

Les documents consultés par les utilisateurs sont analysés de manière à en extraire les termes les plus fréquents. La liste de termes ainsi obtenue permet d'avoir une caractérisation thématique du document.

Pour chaque période de travail la signature thématique d'un utilisateur (profil) est établie en concaténant les listes de termes associées à chaque document consulté par cet utilisateur au cours de cette période.

De nombreux calculs de distance existent pour établir une métrique entre des éléments basés sur de tels profils (voir [2] pour un état de l'art) Nous avons choisit une distance normée, la distance cosine qui donne de bons résultats avec ce type de données.

2.3 Clustering et pistage

Les distances entre chaque couple d'utilisateurs peuvent être rassemblées au sein d'une matrice symétrique.

Un algorithme de clustering calcule les groupes existants sur cette période à partir de cette matrice. Parmi les divers algorithmes existants, nous avons choisit l'algorithme de classification hiérarchique [3]. Son fonctionnement glouton lui assure un temps de calcul raisonnable et une mise en œuvre simplifiée, pour des performances tout à fait correctes.

Nous postulons qu'une communauté peut être considérée comme stable à partir du moment où moins de 50 % de ses membres ont changé d'une période sur l'autre. Au sein d'une même communauté, on autorise donc un renouvellement entre deux périodes consécutives de 50% maximum, que ce soit sur les membres ou sur les thèmes de la communauté. Si plus de la moitié des membres ou des thèmes diffèrent par rapport à la période précédente, alors une nouvelle communauté est déclarée.

3 Visualisation des communautés

Pour observer ces communautés nous avons privilégié certaines caractéristiques dont nous affichons les évolutions au cours des périodes étudiées (voir [4] pour une discussion sur les différentes possibilités).

3.1 Evolution de la population

La population d'une communauté est la première de ces caractéristiques. Elle indique son importance par rapport aux autres communautés, ainsi que sa durée de vie.

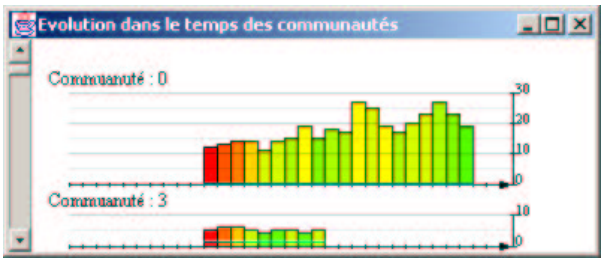

Fig. 1 – Evolution des populations

Sur cette représentation, chaque barre de l'histogramme représente la population de la communauté sur une période donnée. Nous avons utilisé une variation de couleur du rouge au vert sur chaque barre de l'histogramme pour évoquer la similarité entre la composition de la communauté par rapport à sa composition d'origine. Les barres vertes correspondent à un total renouvellement des membres de cette communauté depuis son apparition.

3.2 Renouvellement

Un affichage plus fin du renouvellement permet de démarquer, période par période, les membres qui étaient déjà présent à la période précédente (en gris) des nouveaux venus (en blanc)

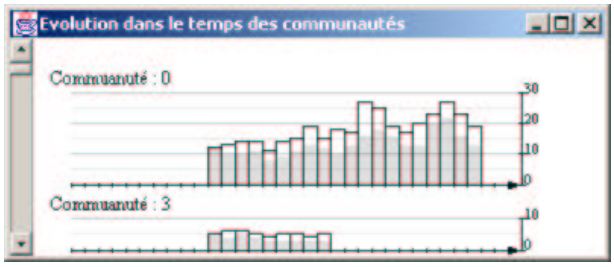

Fig. 2 – Renouvellement de la communauté d'une période à une autre.

Pour une vision plus synthétique, on peut choisir l'affichage du taux de renouvellement en pourcentage : le graphe ci dessous donne le pourcentage de membres anciens à chaque période de la communauté.

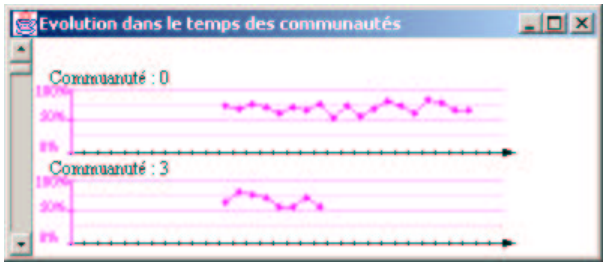

Fig. 3 – Evolution du taux de renouvellement

3.3 Observations

On constate qu'en plus d'informations sur l'évolution dans le temps des communautés vues comme un ensemble homogène, notre outil permet d'avoir une vision plus précise de la dynamique interne de la communauté. En effet, au-delà de l'aspect quantitatif (e.g. évolution de la communauté en nombre), il est possible de percevoir des éléments plus qualitatifs comme la stabilité d'un groupe (i.e le taux d'individus présent d'une période sur l'autre). Comme nous l'avons précisé, une même communauté thématique, basé sur des centres d'intérêts identiques et possédant un nombre de membres identique peut avoir un turn-over de 100 % (i.e. changement total de ses membres). Cette constatation pointe la nécessité de préciser la nature de ce que l'on observe. Une communauté peut aussi bien se décrire par ses membres (e.g une famille) que par ses centres d'intérêts ou d'autres critères. Dans tous ces cas les invariants peuvent être de nature très différente.

4 Conclusion

L'observation de la dynamique interne nécessite une approche particulière de l'ergonomie de la visualisation de manière à mettre en évidence certains caractères de la représentation sans en détruire d'autres. Un des objectifs de notre travail est d'approfondir ces aspects important d'un outil d'observations. A la manière d'un microscope, un bon outil doit être capable de s'introduire dans les détails mais aussi d'avoir une vision d'ensemble. La difficulté dans ce que nous voulons observer c'est que contrairement au microscope dont le champ d'investigation est physique (3 dimensions), notre champ d'investigation peut avoir un nombre de dimensions très important incluant le temps, ce qui rend la tâche plus ardue tant d'un point de vue algorithmique que d'un point de vue conceptuel.

Cet aspect de notre travail est relié à une problématique de fond qui concerne l'identification, la modélisation et l'exploitation des phénomènes d'intelligences collectives [5]

Références

[1]A. Lavallard et L. Lancieri, Outil de représentation des évolutions de communautés d'intérêts, EGC 2004

[2] L. Lee. Measures of distributional similarity. *Proceedings of the 37th ACL*, 1999

[3] P. Ronkainen: Attribute similarity and event sequence similarity in data mining. *Ph.Lic. Thesis, Report C-1998-42, University of Helsinki*, 98 pages,1998.

[4] J. Bertin, *La Graphique et le Traitement graphique de l'information*, Paris, Flammarion, 1977

[5] L. Lancieri: Reusing implicit cooperation, a novel approach in knowledge management; In TripleC (Cognition, Cooperation, communication) international journal 2004.