Découverte de communautés par analyse des usages

Lylia Abrouk, David Gross-Amblard, Damien Leprovost

Découverte de communautés par analyse des usages

Lylia Abrouk, David Gross-Amblard, Damien Leprovost

Laboratoire Le2i-CNRS
Université de Bourgogne, France
{prénom.nom}@u-bourgogne.fr
http://www.u-bourgogne.fr/LE2I

Résumé. Dans les sites Web collaboratifs actuels, un effort de saisie important est demandé aux utilisateurs afin d'identifier la communauté à laquelle ils appartiennent (description du profil personnel, du réseau social, etc.). Dans cet article, nous proposons une méthode de découverte de communautés basée sur les actions des utilisateurs. Elle repose sur une analyse en composantes principales des usages (ACP) et a été validée sur une base de données de préférences filmographiques de grande taille (MovieLens).

1 Introduction

Depuis quelques années, le Web s'est transformé en une plateforme d'échange générique, où tout utilisateur devient un fournisseur de contenu par le biais de technologies comme les commentaires, les blogs et les wikis. Ce nouveau Web collaboratif ou participatif (Web 2.0) comprend des sites populaires comme Myspace[1], Facebook[2] ou Flickr[3], permettant de construire des réseaux sociaux selon ses relations professionnelles ou ses intérêts. Cependant, ces sites exigent de chaque utilisateur une description explicite de son réseau social ou de son profil. De plus, seules les communautés ainsi explicitées sont identifiées.

Or un grand nombre de communautés d'utilisateurs existent de façon implicite dans de nombreux domaines. Par exemple, tout site de musique généraliste rassemble une communauté d'utilisateurs ayant des goûts musicaux variés. Mais cette communauté est en fait composée de sous-communautés potentiellement disjointes, toutes liées à la musique (la communauté des amateurs de musique pop, de musique punk, etc.). Découvrir et identifier précisément ces communautés implicites est un gain pour de nombreux acteurs : le propriétaire du site, les régies publicitaires en ligne et surtout, les utilisateurs du système.

Dans cet article, nous proposons une méthode de détection de communautés. La méthode est générique car elle ne s'appuie que sur un étiquetage des ressources et sur l'utilisation de ces ressources par les utilisateurs (par exemple, tel utilisateur consulte tel fichier musical, étiqueté `rock`). Le cœur de notre méthode est une analyse statistique en composantes principales (ACP (Falissard, 2005)) des étiquettes des ressources manipulées par les utilisateurs. Cette méthode permet de représenter les données originelles (utilisateurs et étiquettes manipulées) dans

[1] http ://www.myspace.com
[2] http ://www.facebook.com
[3] http ://www.flickr.com

un espace de dimension inférieure à celle de l'espace originel, tout en minimisant la perte d'information. La représentation des données dans cet espace de faible dimension en facilite considérablement l'analyse et permet ainsi de regrouper ou d'opposer des communautés.

L'article est organisé de la façon suivante. La section 2 présente notre approche de détection de communautés. Cette approche est validée expérimentalement en section 3 sur une bases de données de préférences filmographiques de grande taille (MovieLens). L'état de l'art est présenté en section 4. Conclusion et perspectives sont présentées en section 5.

2 Modèle

Premières définitions On considère un ensemble d'utilisateurs $U = \{u_1, \ldots, u_n\}$ et un ensemble de ressources R sur un site donné (par exemple des fichiers de musiques, des vidéos, des nouvelles). Nous supposons que les utilisateurs émettent un vote sur un sous-ensemble des ressources du site. Ce vote n'est pas nécessairement explicite et peut être obtenu en se basant sur les usages des utilisateurs (la musique qu'ils sélectionnent, les titres qu'ils achètent, les ressources qu'ils annotent ou recommandent). Les votes sont illustrés par une matrice M : $|U| \times |R|$ définie comme suit, pour un utilisateur $u_i \in U$ et une ressource $r_j \in R$:

$$M(u_i, r_j) = \begin{cases} 1 & \textit{si } u_i \textit{ a de l'intérêt pour } r_j, \\ 0 & \text{sinon.} \end{cases} \quad (1)$$

Cette matrice est mise à jour dynamiquement lorsque de nouveaux utilisateurs, de nouvelles ressources ou de nouveaux usages apparaissent sur le site. Nous supposons également qu'un ensemble de *tags* $T = \{t_1, \ldots, t_m\}$ est défini (par exemple, musique `pop`, `rock`, `punk`, etc.), et que chaque ressource est annotée avec un sous-ensemble de ces tags (sous-ensemble potentiellement vide). Ces annotations proviennent des fournisseurs de ressources, qui peuvent être les utilisateurs eux-même, et peuvent s'enrichir au fur et à mesure. Étant donnés les votes des utilisateurs et ces annotations, nous définissons l'ensemble $A(u_i) \subseteq R$ des ressources intéressant l'utilisateur $u_i \in U$ et l'ensemble $A(u_i, t_j) \subseteq R$, où $t_j \in T$, l'ensemble des ressources intéressant u_i et annotées par le tag t_j.

L'objectif principal de l'approche proposée est de scinder les utilisateurs en communautés distinctes, en se basant sur les groupes de tags qu'ils apprécient. Nous calculons le degré d'appartenance x_{ij} d'un utilisateur u_i à un tag t_j :

$$x_{ij} = \frac{|A(u_i, t_j)|}{|A(u_i)|}. \quad (2)$$

Plus un coefficient x_{ij} est proche de 1, plus l'utilisateur i manipule des tags de type j.

Communautés de tags On cherche ensuite à rassembler les tags similaires, de façon statistique. Pour cela, on utilise la technique de l'analyse en composantes principales (ACP). Dans cette section, nous donnons l'intuition de cette méthode, les détails étant explicités en section 3.

Dans la suite, l'usage d'une ressource portant un tag donné est vu comme la réalisation d'une variable aléatoire représentant ce tag. Les intérêts de chaque utilisateur sont alors autant de réalisations indépendantes des m variables représentant les m tags possibles. L'objectif de

l'ACP est de trouver des combinaisons linéaires des variables représentant les tags pour expliquer au mieux les intérêts des utilisateurs. Ainsi, à chaque utilisateur u_i, nous associons le vecteur de ses degrés d'appartenance à chaque tag, $X_i = (x_{i1}, x_{i2}, ..., x_{im})$. Ce vecteur représente le positionnement de l'utilisateur dans l'espace des tags, et l'ensemble des vecteurs X_i donne ainsi un nuage de points dans l'espace des tags. De la même manière, on peut associer à chaque tag t_j le vecteur V_j, correspondant à ses degrés d'appartenance chez les n utilisateurs : $V_j = (x_{1j}, x_{2j}, .., x_{ij}..., x_{nj})$. Ces nuages de points sont difficiles à analyser, à cause des dimensions considérées (nombre de tags, nombre d'utilisateurs) et de la variabilité des observations. L'analyse en composantes principales va alors :

1. Permettre une projection du nuage de points utilisateurs (initialement exprimés dans un espace de dimension k) sur des plans principaux (de dimension 2) qui reconstituent au mieux la variabilité entre les utilisateurs.
2. Permettre une représentation des variables initiales dans ces plans principaux, la contribution des variables dans la construction des axes principaux n'étant pas la même pour toutes les variables. Par exemple, la figure 1 donne une représentation compacte des rassemblements de tags selon leurs usages.

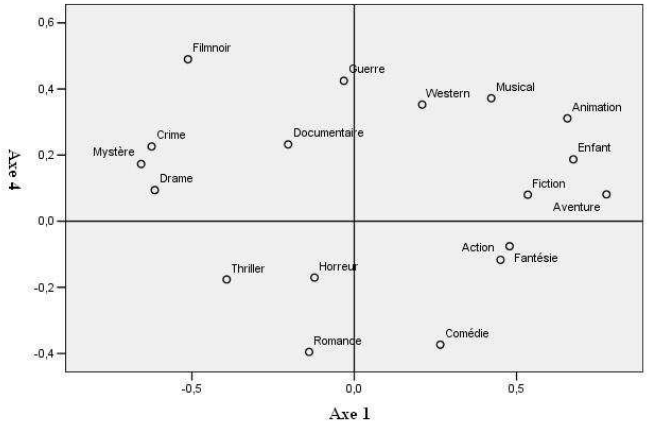

FIG. 1 – *Projection des variables sur deux axes*

Ainsi, des axes explicatifs sont identifiés, en minimisant la perte d'information effectuée lors de cette simplification. La figure 1 représente les variables originales de nos expérimentations sur deux axes significatifs, appelés *composantes principales* (dans cette figure, nommés axes 1 et 4). Cette figure présente la corrélation des variables d'origine avec les composantes principales (une variable est bien représentée sur l'axe si sa corrélation avec la composante principale correspondante est en valeur absolue proche de 1). Selon la composante 1 (Axe 1), on voit que les tags `Animation` et `Enfant` sont très corrélés (corrélation supérieure à 0,6). De même, la composante 4 oppose les tags `Filmnoir`, `Guerre` aux tags `Romance`, `Comédie`.

Notre méthode de rassemblement de tags est alors la suivante : l'ACP fournit les composantes principales pertinentes pour l'analyse des usages. Selon chacune de ces composantes, on ignore les tags situés dans la zone de faible corrélation (corrélation entre $-\alpha$ et $+\alpha$, pour un seuil $\alpha \in]0, 1]$ fixé). Les tags restants, situés dans les zones de forte corrélation (inférieure à $-\alpha$ ou supérieure à $+\alpha$), sont rassemblés dans une même communauté de tags. Par exemple, `Animation` et `Enfant` seront dans une même communauté. L'algorithme 1 résume la méthode.

Algorithme 1 : Découverte

entrées : Vecteurs V_j, seuil de décision α
sorties : Communautés de tags $G_1, \ldots, G_K$

début
  identifier les composantes principales $C = ((c_1, c_2), (c_3, c_4)...)$, expliquant la plus grande proportion de la variabilité des données
  tant que *(il reste des composantes principales* (c, c') *dans* C*)* **faire**
    ignorer les tags non corrélés (|coordonnées selon c et c'| $< \alpha$)
    rassembler dans une même communauté les tags corrélés selon c (|coordonnées selon c| $> \alpha$)
    rassembler dans une autre communauté les tags corrélés selon c' (|coordonnées selon c'| $> \alpha$)
    supprimer ces tags
  fin tant que
fin

Communautés d'utilisateurs Une fois l'ensemble des tags T décomposé en K communautés de tags $G_1, \ldots, G_K$, on en déduit les communautés d'utilisateurs. Pour cela, pour un utilisateur u_i donné, on calcule son degré d'appartenance x'_{ij} à chaque communauté de tag G_j :

$$x'_{ij} = \sum_{t_k \in G_j} x_{ik}.$$

Sa communauté $c(u_i)$ est alors sa communauté de tag majoritaire, c'est à dire l'indice j tel que x'_{ij} soit maximal. Chaque utilisateur est alors associé à ce groupe de tags. Ce groupe aura comme intitulé l'ensemble des tags qui le constituent.

3 Expérimentation

Contexte Nous avons testé la méthode sur la base de films MovieLens[4]. Cette base contient 100 000 votes pour 1 682 films appréciés par 943 utilisateurs. Les films sont évalués par une note entre 1 et 5. Nous avons remplacé ces notes par un vote binaire (les notes supérieures à 2 indiquant un intérêt pour le film). Nous avons construit la matrice M avec l'ensemble des

[4]http ://www.grouplens.org/node/73

utilisateurs U et l'ensemble des films R, et calculé le degré d'appartenance des utilisateurs aux différents tags. Nous présentons les résultats de notre approche sur un ensemble de 18 tags (1 : Aventure, 2 : Enfant, 3 : Animation, 4 : Mystère, 5 : Crime, 6 : Drame, 7 : Fiction, 8 : Filmnoir , 9 : Fantasy, 10 : Musical, 11 : Action, 12 : Thriller, 13 : Romance, 14 : Comédie, 15 : Horreur, 16 : Guerre, 17 : Documentaire, 18 : Western). Le seuil de décision α a été fixé à 0,6 de façon empirique (la sélection automatique de ce seuil n'a pas pu être abordée dans le cadre de ce premier travail.)

Matrice de corrélation La première étape de l'analyse est de vérifier que les données sont factorisables, c'est-à-dire qu'elles sont corrélées entre elles. Pour cela, on examine la matrice de corrélation :

- Si les coefficients de corrélation entre variables sont faibles, il est improbable d'identifier des facteurs communs. On peut éventuellement supprimer les variables qui ont une corrélation faible.
- Un autre paramètre pouvant aider au choix des variables est la qualité de la représentation (*Communalities*) ; QLT_j est le cosinus carré de l'angle formé entre la variable initiale x_j et l'axe principal c.

Le tableau de la figure 2 représente la matrice de corrélation entre une partie des variables initiales et les 6 premières composantes principales.

Tag	1	2	3	4	5	6
Aventure	**,777**	,349	-,272	,081	,037	-,056
Enfant	**,675**	-,231	*,465*	,187	-,145	-,147
Animation	**,657**	-,200	*,391*	,311	-,052	-,218
Mystère	**-,657**	,258	*,367*	,173	-,254	-,057
Crime	**-,624**	,265	,094	,226	,237	-,016
Drame	**-,614**	-,561	-,230	,094	,016	-,112
Fiction	,535	,531	-,252	,080	,249	-,152
Filmnoir	-,512	,066	,209	**,490**	,083	,158
Fantasy	,479	-,108	,197	-,076	,208	-,022
Musical	,422	-,409	,380	,372	-,193	,096
Action	,451	**,746**	-,262	-,117	-,128	,028
Thriller	-,393	**,704**	,314	-,176	-,221	,011
Romance	-,139	**-,685**	-,221	**-,395**	-,231	-,023
Comédie	,265	**-,592**	,161	**-,373**	,225	,242
Horreur	-,122	,424	*,360*	-,170	,369	,179
Guerre	-,032	-,037	**-,633**	**,425**	-,331	-,103
Documentaire	-,204	-,263	-,166	,232	**,639**	-,400
Western	,209	-,105	-,262	,353	,142	**,780**

FIG. 2 – *Corrélation entre les variables et les composantes*

La qualité de la représentation de la variable Action, par exemple, est obtenue en élevant au carré les coefficients de corrélation entre cette variable et les 6 axes principaux, puis en les

sommant :

$$QLT_{\texttt{Action}} = (0,451)^2+(0,746)^2+(0,262)^2+(0,117)^2+(0,128)^2+(0,028)^2 = 0,859.$$

Ainsi pour chaque variable initiale, nous obtenons la variance prise en compte par l'ensemble des facteurs extraits. Plus cette valeur est proche de 1, plus l'ensemble de l'information contenue dans la variable est prise en compte. Il serait par exemple possible de négliger la variable correspondant au tag `Fantasy` en raison de sa faible qualité de représentation (nous l'avons cependant conservée lors de nos expérimentations).

Sélection des composantes principales La deuxième étape consiste à déterminer le nombre de facteurs à retenir. On tient compte :

- des facteurs qui permettent d'extraire une quantité d'information (valeur propre) > 1. Quand on a beaucoup de variables, il y a un grand nombre de facteurs pour lesquels la valeur propre est supérieure à 1. Dans ce cas, on retient beaucoup de facteurs et l'interprétation devient difficile.
- de la distribution des valeurs propres : utilisation du graphique des valeurs propres.

La figure 3 représente la variance expliquée par chaque composante principale (valeur propre). Pour savoir combien de composantes principales utiliser, on recherche une rupture de pente sur le graphique. Cette rupture signifie que l'on passe d'un facteur représentant beaucoup d'information à un facteur en représentant moins. On s'arrête au facteur précédant cette rupture de pente. Dans notre expérimentation, on retient les 6 premières composantes dont la valeur propre est supérieure à 1. Le pourcentage de variance expliquée est de 70%.

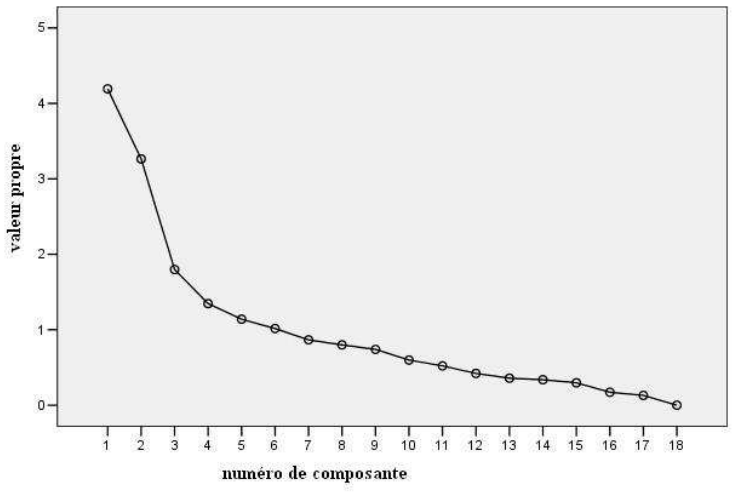

FIG. 3 – *Variance expliquée par chaque composante principale*

Les composantes obtenues ont la structure suivante :

- La 1re composante principale est la combinaison qui totalise la plus grande quantité de variance,
- La 2e composante principale est la combinaison qui totalise la 2ème plus grande quantité de variance. On peut déterminer autant de composantes principales qu'il existe de variables. La valeur propre de la 1re composante principale est 4,192 (soit 23,29% de la variance), celle de la 2e composante est 3,264 (soit 18,13% de la variance),etc. Les composantes principales sont indépendantes les unes des autres.

À partir de la matrice de corrélation, on voit que :

- La 1re composante principale représente essentiellement les variables `Aventure`, `Enfant`, `Animation`, `Mystère`, `Crime` et `Drame`.
- La 2e composante principale représente essentiellement les variables `Action`, `Thriller`, `Romance` et `Comédie`.
- La 3e composante principale représente essentiellement la variable `Guerre` et à un moindre degré les variables `Enfant`, `Animation`, `Mystère` et `Horreur`.
- La 4e composante principale représente essentiellement les variables `Filmnoir`,`Guerre` d'une part, et `Romance`, `Comédie` d'autre part.
- La 5e composante principale représente essentiellement la variable `Documentaire`.
- La 6e composante principale représente essentiellement la variable `Western`.

Interprétation des axes La dernière étape de l'expérimentation est l'interprétation des axes. on donne un sens à un axe à partir des coordonnées des variables. Ce sont les valeurs extrêmes qui concourent à l'élaboration des axes. Les facteurs avec de larges coefficients (en valeur absolue) pour une variable donnée indiquent que ces facteurs sont proches de cette variable. Nous rapprochons les tags par les degrés d'appartenance des utilisateurs à ces tags en nous basant sur les graphiques générés lors de cette étape :

- Le 1er axe (figure 4) oppose les tags `Animation`, `Enfant` et `Aventure` aux tags `Mystère`, `Crime` et `Drame`. Ceci correspond à une interprétation naturelle : les personnes qui aiment le premier groupe de films n'aimant en général pas le second. Deux communautés sont ainsi créées.
- Le 2e axe oppose les films de `Romance` et de `Comédie` aux films `Thriller` et `Action`, en créant ainsi deux nouvelles communautés.
- Le 3e axe (figure 5) oppose les films de `Guerre` aux films étiquetés `Enfant`, d'`Animation` ou de `Mystère`.

Les axes 4, 5 et 6 nous donnent les résultats suivants :

- Le 4e axe oppose les films `Filmnoir` et les films de `Guerre` aux films de `Romance` et de `Comédie`.
- Le 5e axe oppose les films `Documentaire` aux films de `Guerre`.
- Le 6e axe oppose les films `Western` aux films `Documentaire`.

Cette interprétation nous donne 7 groupes de tags, comme indiqué au tableau 1. Les groupes qui sont disjoints sont 1 et 2, 3 et 4, 4 et 6 et enfin 6 et 7. Les utilisateurs sont regroupés en fonction de ces communautés de tags. Les tags qui ne sont pas pris en compte par les axes sont expliqués par leur faible occurrence : par exemple le tag `Fantasy` n'est utilisé que 22 fois sur toute la collection des 1682 films.

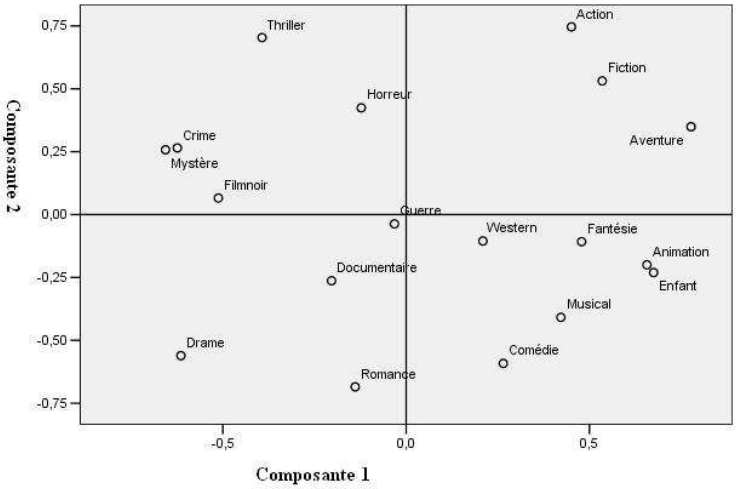

FIG. 4 – *Composantes 1 et 2*

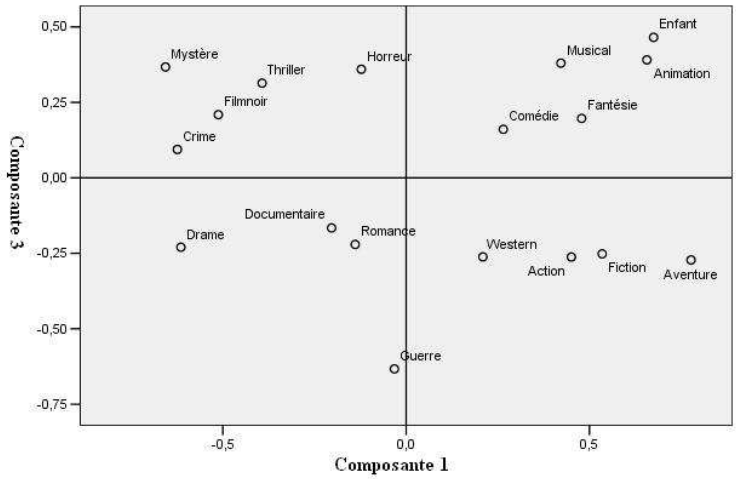

FIG. 5 – *Composantes 1 et 3*

communauté	tags associés
1	Aventure, Enfant, Animation
2	Mystère, Crime, Drame
3	Action, Thriller
4	Romance, Comédie
5	Western
6	Filmnoir, Guerre
7	Documentaire

TAB. 1 – *Communautés de tags*

4 État de l'art

Depuis les débuts du Web jusqu'à aujourd'hui, la recherche de communautés implicites a fortement évolué. De nombreux travaux, envisagent divers aspects des réseaux sociaux et des communautés, selon que l'on considère une communauté comme un ensemble de documents relatif à une thématique, ou comme un ensemble de personnes partageant un intérêt pour une thématique.

Découverte des communautés Web Dès les premiers travaux sur la reconnaissance des communautés sur le Web (par exemple Gibson et al. (1998)), le lien hypertexte est utilisé comme base de raisonnement. L'apport majeur en la matière est l'algorithme HITS de Kleinberg (1998), définissant les notions d'autorités et de *hubs*, structurant une communauté autour d'un sujet donné. Imafuji et Kitsuregawa (2002) concluent à l'appartenance d'une page à une communauté si cette page est plus majoritairement référencée depuis l'intérieur de la communauté que depuis son extérieur. Ils utilisent un algorithme de flot maximum afin d'isoler les noeuds faisant partie d'une même communauté, en se basant sur l'algorithme proposé par Flake et al. (2000). Dourisboure et al. (2007) identifient au sein d'un graphe du Web les communautés comme autant de sous-graphes denses et bipartis au sein de ce graphe. Le graphe biparti représente d'une part les centres d'intérêt de la communauté (les autorités selon HITS) et d'autre part ceux qui citent la communauté (les *hubs*). Cette méthode permet de mettre en évidence les éventuels partages des mêmes centres d'intérêt par plusieurs communautés d'acteurs, ou au contraire le partage de mêmes acteurs par plusieurs centres d'intérêt des communautés. Ces approches fournissent une analyse avancée des liaisons entre les différentes pages structurant une communauté thématique, mais ne permettent pas en revanche de rapprocher des utilisateurs de par leurs intérêts ou activités : le partage de lien hypertexte n'étant plus nécessairement la base de l'activité communautaire dans les échanges sociaux du Web collaboratif (évaluation de contenu par l'utilisateur, apposition de tags, ...).

Interprétations des tags utilisateurs Les systèmes de recommandations proposent à l'utilisateur un lot de ressources en corrélation avec son profil ou son activité. Firan et al. (2007) proposent un algorithme de recommandation basés sur les tags des utilisateurs. Ils prennent pour

exemple l'utilisation des tags sur le site de musique Last.fm[5], où les pistes musicales sont filtrées en fonction des classements (votes) personnels de l'utilisateur. Cette méthode se heurte au problème de l'initialisation (*cold start*), les nouveaux utilisateurs recevant d'abord des recommandations peu pertinentes. Une solution hybride (basée sur l'aspect collaboratif, mais aussi sur le contenu) proposée par Yoshii et al. (2006) utilise un modèle probabiliste pour intégrer les votes utilisateurs et le contenu des données, en utilisant un réseau bayésien pour améliorer les méthodes classiques. Permettant un positionnement pertinent de l'utilisateur par rapport aux tags du système, ces solutions ne permettent pas de tenir compte des possibles similarités entre tags. La mise en lumière des tags similaires ou antagonistes que propose notre solution permet d'affiner ce positionnement de l'utilisateur.

Distances sémantiques Cattuto et al. (2008) présentent une autre approche statistique pour évaluer les distances sémantiques. Validée sur les données du site `del.icio.us`[6], site sur lequel il existe une structure communautaire, les auteurs utilisent l'annotation des données pour construire un réseau pondéré de ressources. Dans ce contexte, la similarité entre les ressources est proportionnelle au chevauchement de leurs jeux de tags. Pour prendre en compte la représentativité des tags, la méthode TF-IDF est utilisée. Les auteurs proposent de détecter les communautés d'utilisateurs par les similarités de leurs tags. Ils utilisent le coefficient de corrélation de Pearson comme mesure de similarité, puis appliquent des méthodes de partitionnement. À la différence de notre méthode, ils ne réduisent pas le nombre de tags manipulés, qui risque d'être extrêmement grand.

Systèmes de recommandation Le rapprochement de tag est également abordé dans les systèmes de recommandation. Dans leur définition du système *Socialranking*, Zanardi et Capra (2008) procèdent à un enrichissement de requête basé notamment sur la similarité des tags, fondée sur leurs apparitions communes sur des ressources différentes. Une autre approche, proposée par Hotho et al. (2006) sous le nom *FolkRank* et utilisant à nouveau de la théorie des graphes, consiste à utiliser *PageRank* pour modéliser les relations entres les ressources, les utilisateurs et les tags. Cette approche, qui permet d'exploiter d'avantage les relations éparses, est également explorée par Bertier et al. (2009) : dans le cadre de *Gossple*, les auteurs utilisent la probabilité de passer d'un tag à un autre comme indicateur de leur similarité.

5 Conclusion et perspectives

Dans cet article, nous avons présenté une méthode de découvertes de communautés d'utilisateurs par observation des usages, basée sur la technique de l'ACP. Une prochaine étape consiste en l'automatisation complète de la méthode, en particulier par l'estimation fine et automatique des seuils de sélection à utiliser, ainsi que la comparaison avec d'autres méthodes statistiques.

[5] http ://www.lastfm.com
[6] http ://delicious.com

Remerciements

Ce travail est partiellement financé par l'ANR Contenu & Interaction Neuma 2008-2011[7] et le projet CheckSem[8].

Références

Bertier, M., R. Guerraoui, V. Leroy, et A.-M. Kermarrec (2009). Toward personalized query expansion. In *SNS '09 : Proceedings of the Second ACM EuroSys Workshop on Social Network Systems*, New York, NY, USA, pp. 7–12. ACM.

Cattuto, C., A. Baldassarri, V. D. P. Servedio, et V. Loreto (2008). Emergent community structure in social tagging systems. *Advances in Complex Systems (ACS) 11*(04), 597–608.

Dourisboure, Y., F. Geraci, et M. Pellegrini (2007). Extraction and classification of dense communities in the Web. In *WWW'07 : Proceedings of the 16th international conference on World Wide Web*, New York, NY, USA, pp. 461–470. ACM.

Falissard, B. (2005). *Comprendre et utiliser les statistiques dans les sciences de la vie*. Masson, Paris.

Firan, C. S., W. Nejdl, et R. Paiu (2007). The benefit of using tag-based profiles. In *LA-WEB '07 : Proceedings of the 2007 Latin American Web Conference*, Washington, DC, USA, pp. 32–41. IEEE Computer Society.

Flake, G. W., S. Lawrence, et C. L. Giles (2000). Efficient identification of Web communities. In *KDD'00 : Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, New York, NY, USA, pp. 150–160. ACM.

Gibson, D., J. Kleinberg, et P. Raghavan (1998). Inferring Web communities from link topology. In *HYPERTEXT'98 : Proceedings of the ninth ACM conference on Hypertext and hypermedia : links, objects, time and space—structure in hypermedia systems*, New York, NY, USA, pp. 225–234. ACM.

Hotho, A., R. Jäschke, C. Schmitz, et G. Stumme (2006). Folkrank : A ranking algorithm for folksonomies. In *Proc. FGIR 2006*.

Imafuji, N. et M. Kitsuregawa (2002). Effects of maximum flow algorithm on identifying Web community. In *WIDM'02 : Proceedings of the 4th international workshop on Web information and data management*, New York, NY, USA, pp. 43–48. ACM.

Kleinberg, J. M. (1998). Authoritative sources in a hyperlinked environment. In *SODA'98 : Proceedings of the ninth annual ACM-SIAM symposium on Discrete algorithms*, Philadelphia, PA, USA, pp. 668–677. Society for Industrial and Applied Mathematics.

Yoshii, K., M. Goto, K. Komatani, T. Ogata, et H. G. Okuno (2006). Hybrid collaborative and content-based music recommendation using probabilistic model with latent user preferences. In *ISMIR'06 : 7th International Conference on Music Information Retrieval*, pp. 296–301.

[7] http ://neuma.irpmf-cnrs.fr
[8] http ://iutdijon.u-bourgogne.fr/checksem

Zanardi, V. et L. Capra (2008). Social ranking : uncovering relevant content using tag-based recommender systems. In *RecSys '08 : Proceedings of the 2008 ACM conference on Recommender systems*, New York, NY, USA, pp. 51–58. ACM.

Summary

Most of the existing social network systems require from their users an explicit statement of their friendship relations. In this paper we focus on implicit communities of users and present an approach to automatically detect communities of Web users, based on user's resource manipulations. Our proposal relies on the Principal component analysis (PCA) method and is assessed on a large movie data set.