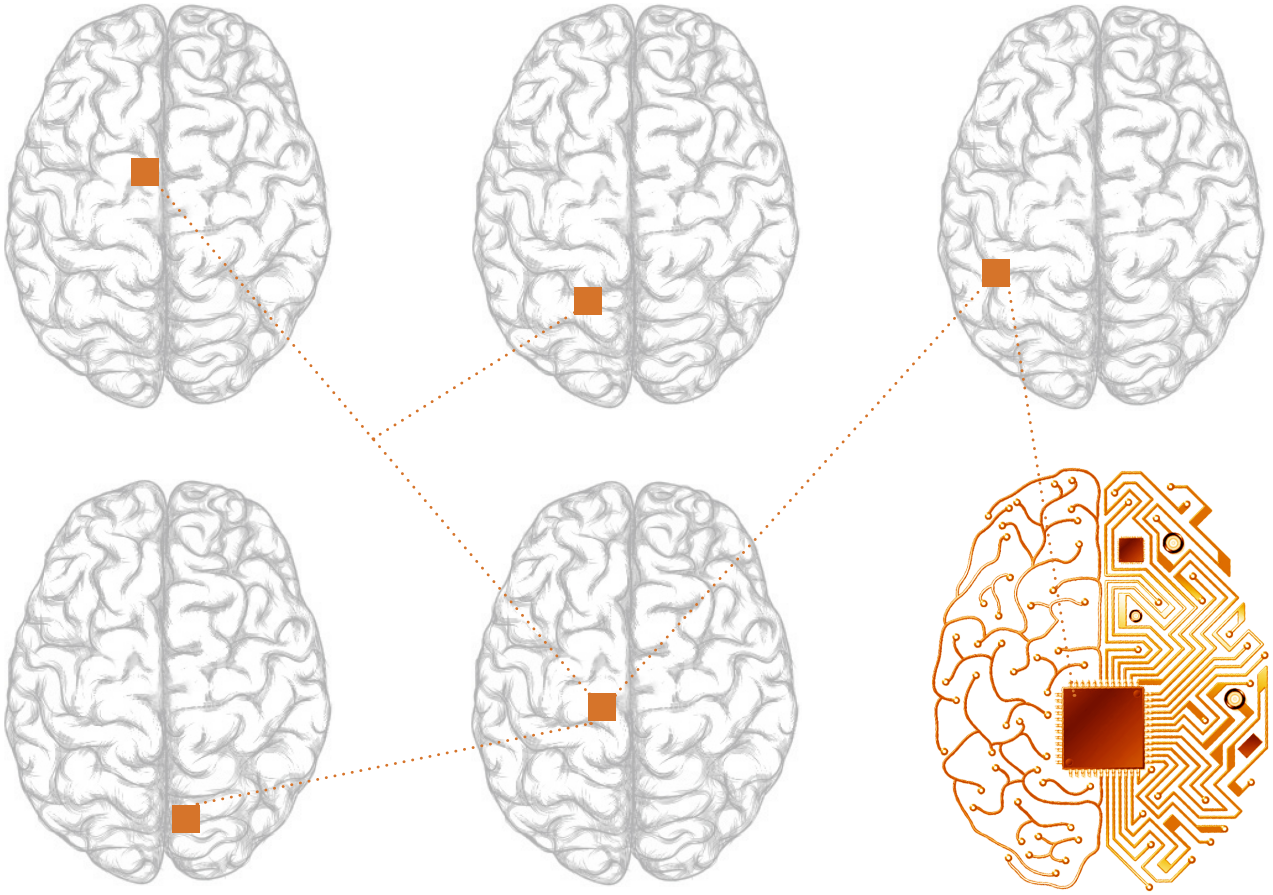


PFIA 2015

Plate-forme Intelligence Artificielle

Rennes

Actes RS et IA

Présidents CP : Tassadit Bouadi & Arnaud Martin

Inria
INVENTEURS DU MONDE NUMÉRIQUE

AfIA
Association française
pour l'Intelligence Artificielle

Réseaux Sociaux et Intelligence Artificielle

Présentation de la conférence

Ces dernières années ont été marquées par l'explosion des réseaux sociaux sur Internet, renouvelant l'intérêt de la communauté scientifique non seulement en sciences sociales mais également en informatique pour l'analyse et la fouille de graphes. Aujourd'hui, les graphes utilisés pour étudier ces réseaux sont de très grande taille tant par le nombre de n$\frac{1}{2}$uds associés aux individus que par celui des arêtes qui décrivent leurs relations ou leurs interactions. Les besoins d'analyse ont également évolué, requérant le développement d'algorithmes et d'outils plus puissants pouvant algorithmiquement passer à l'échelle, tant en masse de données qu'en débit.

L'apprentissage artificiel, discipline au croisement de l'informatique et des statistiques, tente de répondre à ces nouveaux défis posés par les innombrables applications autour des réseaux sociaux : passage à l'échelle (faire face à la volumétrie), prise de décisions dans des environnements évolutifs et complexes, prise en compte de la sémantique des communications (taxonomies, ontologies, controverses, analyse des sentiments), augmentation de l'hétérogénéité des sources de données, etc.

L'objectif de cet atelier est de permettre aux chercheurs, industriels et étudiants de positionner les problématiques du domaine et d'identifier les avancées récentes et les problèmes ouverts.

Cet atelier vise à réunir des contributions scientifiques, et à échanger des points de vue et retours d'expériences, sur le thème de l'utilisation de techniques d'apprentissage artificiel pour l'analyse de réseaux sociaux.

Un exposé invité sur le thème de la validation de détection de communautés complétera le programme afin de stimuler l'échange sur les domaines intéressants de l'atelier.

En plus des points mentionnés précédemment, les thèmes de cet atelier sont les suivants (liste non limitative) :

— Fouille de grands graphes sociaux (clustering, classification, recherche de motifs fréquents)
— Recommandation sociale
— Prédiction de liens dans les réseaux sociaux
— Détection du Spam social
— Qualification des données dans les réseaux sociaux (fiabilité, imprécision, incertitude, etc).
— Analyse de données sociales et extraction de connaissances
— Détection et évolution de communautés
— Analyse des réseaux sociaux dynamiques
— Analyse des sentiments et fouille d'opinions
— Réseaux sociaux et ingénierie des connaissances

Conférencier Invité : Christine Largeron (Professeur à l'Université Jean Monnet, Saint-Étienne, France)

— Thème de la présentation : "Community detection validation using networks generation"

Comité de Programme

Présidents du comité de programme

Tassadit Bouadi, Université Rennes 1
Arnaud Martin, Université Rennes 1

Membres du comité de programme

— Frédéric Amblard (IRIT, Université Toulouse 1 Capitole)
— Hanene Azzag (LIPN, Université Paris 13)
— Tassadit Bouadi (IRISA, Université Rennes 1)
— Mohand Boughanem (IRIT, Université Paul Sabatier)
— Guillaume Cleuziou (LIFO, Université d'Orléans)
— Cécile Favre (ERIC, Université Lyon 2)
— Chihab Hanachi (IRIT, Université Toulouse 1 Capitole)

— Vincent Labatut (LIA, Université d'Avignon et des Pays du Vaucluse)
— Christine Largeron (LHC, Université Jean Monnet)
— Matthieu Latapy (LIP6-CNRS, Université Pierre et Marie Curie)
— Vincent Leroy (LIG, Université de Grenoble)
— Maria Malek, (LARIS, EISTI - Campus de Cergy)
— Arnaud Martin (IRISA, Université de Rennes 1)
— Rokia Missaoui (LARIM, Université du Québec en Outaouais)
— Jacky Montmain (LGI2P, École des Mines d'Alès)
— Filipo Perotto (IRIT, Université Toulouse 1 Capitole)
— Chantal Reynaud (LRI, Université Paris-Sud)
— Camille Roth (CNRS)
— Kavé Salamatian (LISTIC, Université de Savoie Mont Blanc)
— Henry Soldano (LIPN, Université Paris-Nord)
— Julien Subercaze (LHC, Université Jean Monnet)
— Tanguy Urvoy (Orange labs)
— Julien Velcin (ERIC, Université Lyon 2)

Table des matières

Un nouvel algorithme de propagation de labels avec barrages

Jean−Philippe Attal* — **Maria Malek****

* *EISTI: Ecole Internationale des Sciences du Traitement de l'Information, laboratoire Quartz, Cergy-France 95000*
ETIS-ENSEA Universite de Cergy-Pontoise CNRS France
Email: jal@eisti.eu
** *EISTI: Ecole Internationale des Sciences du Traitement de l'Information, laboratoire Quartz, Cergy-France 95000*
Email: mma@eisti.eu

RÉSUMÉ. *La propagation de labels est l'une des méthodes les plus rapides pour la détection de communautés, de complexité quasi−linéaire en terme d'arêtes. Il s'agit d'une méthode locale où chaque nœud possède son propre label qui change par interaction avec son voisinage. Malheureusement, cette méthode présente deux inconvénients majeurs. Le premier est qu'une mauvaise propagation peut mener à de trop grandes communautés (le problème des communautés monstrueuses). Le second inconvénient est l'instabilité de la méthode, ne donnant que très rarement le même résultat après chaque lancement. Dans cet article, nous proposons des algorithmes et une étude portant sur la propagation de labels en plaçant des barrages sur certaines arêtes dans le but d'éviter de mauvaises propagations. Puis nous appliquons une méthode d'ensemble learning basée sur l'alimentation d'une matrice de fréquence de co-apparition dans le but de stabiliser la propagation de labels. Nous exposons de nouveaux algorithmes d'ensemble learning pour la détection de communautés.*

ABSTRACT.

MOTS-CLÉS : *détection de communautés*$_1$, *propagation de labels*$_2$, *barrages*$_3$, *ensemble learning*$_4$.

KEYWORDS:

1. Introduction

La plupart des réseaux représentant des systèmes réels montrent des caractéristiques propres comme des groupes de noeuds fortement liés entre eux (que l'on appelle des communautés) et peu avec le reste du graphe. Une étude comparative a été effectuée par Fortunato et al. (Fortunato, 2010).

Dans cet article, nous exposons un nouvel algorithme de détection de communautés (et ses variantes), basé sur l'information globale du graphe dans le but d'aider une méthode locale à détecter les communautés.
Nous verrons qu'en limitant la propagation de labels par des barrages, nous pourrons éviter l'obtention de trop grandes communautés. Une méthode basée sur la fréquence d'apparition des noeuds dans une même communauté permettra de détecter des coeurs (ensemble de noeuds fortement liés) et des communautés. Nous expliquerons dans la section 2 la propagation de labels et la détection de coeurs, puis dans la section 3, notre approche liant barrages et stabilisation, puis nos expérimentations dans la section 4, et nous concluerons dans la section 5 avec nos futures perspectives .

2. L'approche par propagation de labels et détection de coeurs

La méthode de propagation de labels (Raghavan *et al.*, 2007) est basée sur la transmission d'informations d'un nœud à ses voisins. L'information étant un label. Un état d'équilibre est atteint lorsque chaque noeud a son label égal à celui de la majorité de ses voisins. Cet algorithme étant très instable, il serait souhaitable d'avoir une méthode de stabilisation. Pour ce faire, une méthode consiste à lancer plusieurs fois l'algorithme indéterministe et à considérer les nœuds qui apparaissent le plus souvent ensemble dans une même communauté. On appelle ces nœuds dont la fréquence d'apparition est très forte, des *cœurs*. Nous proposons d'utiliser la méthode de Seifi et al. (Seifi *et al.*, 2013). Nous nommerons cette méthode le *stabilisateur* Une étude sur le clustering d'ensemble pour les graphes a été effectuée par Mikaël Ovelgönne (Ovelgönne et Geyer-Schulz, 2012).

3. Approche proposée

La propagation de labels avec barrages a pour objectif de détecter des communautés, utilisant à la fois l'information globale topologique du graphe et une méthode locale pour trouver les communautés. La propagation de labels souffrant d'instabilité et du fait de produire de trop grandes communautés, nous proposons une méthode pour résoudre ces deux problèmes en interdisant à certaines arêtes de propager un label tout en effectuant une méthode de détection de cœurs en lançant plusieurs fois l'algorithme indéterministe.

Les mesures issues de l'analyse des réseaux sociaux peuvent se révéler très utiles pour connaître l'importance de certaines arêtes ou nœuds au sein d'un graphe. Nous utilisons ici la mesure d'intermédiarité (Girvan et Newman, 2002) qui nous permettra de placer des barrages lors de l'exécution de la propagation de labels.

Nous exposons l'algorithme de propagation de labels avec barrages (PLAB) 1 :

Algorithme 1 PLAB

Paramètres : Un graphe $G = (V, E)$, un réel β
Sortie : Une partition $P = \{P_1, ..., P_{\mathcal{C}}\}$ (communautés de G)

1: Calcul de la centralité d'intermédiarité de G.
2: Selectionner β pourcentage des arêtes ayant les plus grandes valeurs d'intermédiarité et mettre des barrages.
3: Lancer la propagation de labels standard.
4: **Retourner** Une partition $P = \{P_1, ..., P_{\mathcal{C}}\}$.

La complexité de la centralité d'intermédiarité est en $\Theta(n^2)$ (Fortunato, 2010), donc la complexité de PLAB est en $\Theta(n^2 + m - \beta \times m)$.

Cependant, cet algorithme présente deux inconvénients, le choix du nombre de barrages, et l'instabilité réduite, mais toujours présente. Pour remédier à ces problèmes nous proposons deux algorithmes basés sur l'ensemble learning. Le premier est basé sur l'optimisation d'une fonction de qualité qui peut par exemple être la modularité ou la conductance 2.

Algorithme 2 MPLBS

Paramètres : Un graphe $G = (V, E)$, un seuil α, $\mathcal{N}$ le nombre d'essais, Δ le pas
Sortie : communautés de G

1: Calcul de la centralité d'intermédiarité de G.
2: **Pour** $i = 0$ à 1 avec un pas de Δ **Faire**
3: Mettre des barrages sur les $\Delta \times |E|$ (β) arêtes ayant les plus grandes valeurs d'intermédiarité
4: Calcul du *stabilisateur* avec α en utilisant la propagation de labels
5: **Fin Pour**
6: **Retourner** La partition avec le meilleur score de la fonction de qualité choisie par l'utilisateur : $P = \{P_1, ..., P_{\mathcal{C}}\}$.

L'algorithme nécessite comme paramétrisation α, β et $\mathcal{N}$.
La complexité de l'algorithme est en $\Theta(n^2) + \Theta(\frac{1}{\Delta} \times \mathcal{N} \times k \times (m - \beta m))$.

Le second algorithme fait varier β en alimentant une seule matrice de co−appartenance, puis détecte les composantes connexes. L'idée est qu'alimenter une matrice de fréquence de co-appartenance par une même méthode qui ne produit pas toujours de bons résultats ne peut pas être satisfaisante. Cependant, alimenter une matrice, avec différentes méthodes, ayant différents niveaux de barrage pourrait améliorer la qualité des composantes connexes trouvées pour la détection de communauté. Nous

exposons l'algorithme de propagation de labels avec barrages utilisant la stabilisation (PLBS) 3 :

Algorithme 3 PLBS

Paramètres : Un graphe $G = (V, E)$, un seuil α, $\mathcal{N}$ le nombre d'essais, Δ le pas
Sortie : communautés de G

1: Calcul de la centralité d'intermédiarité de G.
2: Allocation d'une matrice de fréquence de co-appartenance vide
3: **Pour** $i = 0$ à 1 avec un pas Δ **Faire**
4: Mettre des barrages sur les $\Delta \times |E|$ (β) arêtes ayant les plus grandes valeurs d'intermédiarité .
5: Lancer $\mathcal{N}$ fois la propagation de labels avec un nombre différent de barrages.
6: Remplir la matrice de fréquence de co-appartenance avec les résultats des différentes propagations de labels.
7: **Fin Pour**
8: Créer un nouveau graphe $G' = (V, E')$ en partant de $P_{ij}^{\mathcal{N}}$ avec des arêtes dont la pondération est égale ou supérieure à α
9: Creér une partition P en considérant les $\mathcal{C}$ composantes connexes.
10: **Retourner** La partition $P = \{P_1, ..., P_{\mathcal{C}}\}$.

La complexité de l'algorithme est en $\Theta(n^2) + \Theta(\frac{1}{\Delta} \times \mathcal{N} \times k \times (m - \beta m))$.

4. Expérimentations et analyse

Nous avons expérimenté nos algorithmes sur des réseaux connus de la littérature, un réseau footballistique (Girvan et Newman, 2002) et un réseau de dauphins (Lusseau *et al.*, 2003). Nous prenons $\mathcal{N} = 100$. Nous avons effectué une étude comparative avec des algorithmes issus de la littérature, comme la propagation de labels, des variantes comme Lovro et al. (Šubelj et Bajec, 2011) LPA, DPA, Leung et al. (Leung *et al.*, 2009), le spin model (Ronhovde et Nussinov, 2010), Girvan Newman (GN) (Girvan et Newman, 2002)et enfin Louvain (Blondel *et al.*, 2008).
Pour établir les comparaisons, nous utilisons des mesures supervisées (lorsque nous connaissons les vraies communautés) telles que l'information mutuelle normalisée (NMI) (Ana et Jain, 2003), la pureté, l'index de Rand ajusté (ARI) mais également non supervisées telle que la modularité Q (Newman et Girvan, 2004). et la conductance Φ (Kannan *et al.*, 2004)

Réseaux	$\|V\|$ et $\|E\|$	DM	Diamètre	CCM	Réseaux	$\|V\|$ et $\|E\|$	DDM	Diamètre	CCM
Foot	$115 \setminus 615$	10.6957	4.0	0.4072	Dol	$62 \setminus 159$	5.129	8.0	0.3088

Tableau 1. *Caractéristiques avec DM, le degré moyen et CCM, le coefficient de clustering moyen*

Experiences avec d'autres algorithmes													
Algorithms	Q	Φ	NMI	ARI	Pureté	#	*Algorithms*	Q	Φ	NMI	ARI	Pureté	#
Foot #11							*Dol* #2						
Louvain	0.6021	0.2778	0.855	0.7277	0.9217	9	Louvain	0.5185	0.2451	0.5109	0.3274	0.9677	5
GN	0.6005	0.290	0.8788	0.7781	0.8347	10	GN	0.5193	0.2001	0.5541	0.3949	0.9838	5
Spin	0.6052	0.293	0.8922	0.8165	0.86956	10	Spin	0.5285	0.2339	0.5864	0.3734	1.0	5
Leung	0.5689	0.353	0.900	0.8361	0.9217	13	Leung	0.5189	0.2708	0.4811	0.2908	0.9677	6
LPA	0.5946	0.2974	0.8865	0.7787	0.8544	10.39*	LPA	0.4831	0.1720	0.6226	0.5024	0.9790	3.85*
DPA	0.606		0.897				DPA	0.529		0.774			
LICOD	0.49		0.83	0.69		16	LICOD	0.35		0.41	0.32		2
PLBS	0.5822	0.3377	**0.9371**	**0.9154**	0.9652	17	**PLBS**	0.3237	0.1682	0.578	0.6689	0.9677	8
PLAB	0.5985	0.3167	0.9289	0.9004	0.9391	13	**PLAB**	0.3761	0.0549	**0.9429**	**0.9563**	1.0	3
MPLBS	0.6019	0.3102	0.9269	0.8893	0.9304	12	**MPLBS**	0.4569	0.0431	0.5979	0.6671	0.8476	2

Tableau 2. *Résultats comparatifs avec d'autres algorithmes (# signifie le nombre de communautés et $*$ signifie que le résultat est une moyenne sur* 100 *lancements car instable.)*

Foot #11	PLAB	PLBS	MPLBS	*Dol* #2	PLAB	PLBS	MPLBS
α	$\{0.6\}$	$\{0.5\}$	$\{0.5\}$	α	$\{0.6\}$	$\{0.5\}$	$\{0.3-0.5\}$
β	$\{0.05\}$	$-$	$\{0.3\}$	β	$\{0.05\}$	$-$	$\{0.3\}$

Tableau 3. *Paramétrisation de nos algorithmes*

Les observations montrent que les algorithmes MPLBS et PLAB proposés sont très compétitifs par rapport à ceux issus de la littérature, donnant de très bonnes valeurs pour le NMI et le ARI, ainsi que notre version finale PLBS. En effet, PLBS pour le réseau de football obtient un NMI de 0.9371 et un ARI de 0.9154.
Nos expérimentations ont également montré qu'alimenter une matrice d'appartenance avec différents barrages pouvait donner de très bons résultats comme le montre le PBLS. En effet, selon les résultats des mesures supervisées, PLBS dont l'alimentation de la matrice de fréquence est différente de celle du MPLBS montre qu'elle donne de meilleurs résultats. En effet, imposer des barrages limite la propagation de labels et évite le fait de tomber dans de trop grandes communautés. C'est ce que montrent les résultats de PLAB qui nécessitent cependant de tester β.Pour la paramétrisation, nous avons choisi $\alpha = 0.5$ pour tous les réseaux de PLBS. Un α trop fort donnerait de trop nombreuses communautés, et trop faible, risquerait de ne donner qu'une communauté. Le MPLBS utilisant la modularité, nous retourne la paramétrisation maximisant cette dernière.

5. Conclusion et perspectives

Dans cet article, nous avons exposé de nouveaux algorithmes basés sur la propagation de labels. Nos expérimentations ont montré que notre méthode hybride, liant propagation de labels et barrages issus de l'intermédiarité des arêtes permettait l'obtention de résultats satisfaisants. Nous avons également observé qu'alimenter une matrice de fréquences de co-apparition en faisant varier le nombre de barrages permettait

d'augmenter la qualité des communautés obtenues. Nous adaptons actuellement ces algorithmes sur les grands graphes ayant plusieurs milliards d'arêtes. C'est en ce sens que nous développons une solution Hadoop et Spark qui nous permettra le développement de la propagation de labels en parallèle. Dans le cas de PLAB, une étude de l'estimation de α et β sera appronfondie.

Remerciements

Ce projet a été financé par Square Predict, projet "investissement d'avenir" pour le Big Data et les assurances.

6. Bibliographie

Ana L., Jain A. K., « Robust data clustering », *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*, vol. 2, IEEE, p. II-128, 2003.

Blondel V., Guillaume J., Lambiotte R., Mech E., « Fast unfolding of communities in large networks », *J. Stat. Mech*p. P10008, 2008.

Fortunato S., « Community detection in graphs », *Physics Reports*, vol. 486, n° 3, p. 75-174, 2010.

Girvan M., Newman M. E. J., « Community structure in social and biological networks », *Proceedings of the National Academy of Sciences*, vol. 99, n° 12, p. 7821-7826, 2002.

Kannan R., Vempala S., Vetta A., « On clusterings : Good, bad and spectral », *Journal of the ACM (JACM)*, vol. 51, n° 3, p. 497-515, 2004.

Leung I. X., Hui P., Lio P., Crowcroft J., « Towards real-time community detection in large networks », *Physical Review E*, vol. 79, n° 6, p. 066107, 2009.

Lusseau D., Schneider K., Boisseau O. J., Haase P., Slooten E., Dawson S. M., « The bottlenose dolphin community of Doubtful Sound features a large proportion of long-lasting associations », *Behavioral Ecology and Sociobiology*, vol. 54, n° 4, p. 396-405, 2003.

Newman M. E. J., Girvan M., « Finding and evaluating community structure in networks », *Phys. Rev. E*, vol. 69, n° 2, p. 026113, February, 2004.

Ovelgönne M., Geyer-Schulz A., « An ensemble learning strategy for graph clustering. », *Graph Partitioning and Graph Clustering*, vol. 588, p. 187, 2012.

Raghavan U. N., Albert R., Kumara S., « Near linear time algorithm to detect community structures in large-scale networks », *Physical Review E*, vol. 76, n° 3, p. 036106, 2007.

Ronhovde P., Nussinov Z., « Local resolution-limit-free Potts model for community detection », *Phys. Rev. E*, vol. 81, p. 046114, Apr, 2010.

Seifi M., Junier I., Rouquier J.-B., Iskrov S., Guillaume J.-L., « Stable community cores in complex networks », *Complex Networks*, Springer, p. 87-98, 2013.

Šubelj L., Bajec M., « Unfolding communities in large complex networks : Combining defensive and offensive label propagation for core extraction », *Physical Review E*, vol. 83, n° 3, p. 036103, 2011.

Détection des experts dans un cadre incertain

Dorra Attiaoui [1], **Arnaud Martin** [2], **Boutheina Ben Yaghlane** [3]

1. *LARODEC, ISG Tunis, Université de Tunis, Tunisie*
   DRUID, IRISA, Université de Rennes 1, France
   attiaoui.dorra@gmail.com
2. *DRUID, IRISA, Université de Rennes 1, France*
   Arnaud.Martin@univ-rennes1.fr
3. *LARODEC, IHEC Carthage, Université de Carthage, Tunisie*
   boutheina.yaghlane@ihec.rnu.tn

RÉSUMÉ. *Dans cet article nous proposons une modélisation statistique des utilisateurs dans les sites communautaires de questions réponses fondée sur les théories de l'incertain. Nous utilisons une méthode de prise de décision crédibiliste fondée sur la combinaison des informations relatives à chaque type d'utilisateur. Cette approche utilise les probabilités pignistiques pour identifier chaque classe et ainsi permettre une détection des experts selon un thème spécifique.*

ABSTRACT. *In this paper we propose a statistical model for representing users in social networks and more precisely in Question Answering Communities based on uncertainty theories. The theory of fuzzy sets and the theory of belief functions allow to represent and manage uncertainty. We use their mathematical background to combine users informations and then apply an evidential decision method based on the pignistic transformation in order to classify users and detect the experts given a specific topic.*

MOTS-CLÉS : *Experts, théories de l'incertain, théorie des fonctions de croyance, classification*

KEYWORDS: *Experts,uncertainty theories, theory of belief functions, combination, classification*

DOI:10.3166/PFIA.1.1-?? © 2015 Lavoisier

1. Introduction

Dans une époque où le virtuel prime sur notre société, les gens ont acquis de nouveaux réflexes pour obtenir et consommer de l'information sur la toile. Entre sites web spécialisés, réseaux sociaux, l'utilisateur se retrouve souvent confronté à des informations qui peuvent tout aussi bien être exactes, mais parfois contradictoires, voire fausses. La récente émergence des sites communautaires de questions réponses, les a rendus très populaires auprès des internautes. Citons à titre d'exemple des sites comme Yahoo!Answers, Quora, Stackoverflow ou encore Comment ça marche. Organisés selon des thèmes et des sujets bien définis, ils permettent aux utilisateurs de poster des questions et d'y répondre. Ouverts à tous, nous nous trouvons ainsi confronté à des réponses émanant d'experts, de personnes peu formées, voire des trolls.

(Bouguessa *et al*., 2008) ont proposé que les personnes qui ont le plus de connaissances dans un domaine sont celles qui ont donné des réponses étant élues comme les meilleures dans un domaine spécifique. Cependant, (Gjergji *et al*., 2011) ont identifié trois niveaux d'incertitude dans ces sites, le premier est lié à l'extraction et l'intégration des données, le second aux sources d'information et finalement aux informations elles-mêmes. Ainsi, identifier les experts parmi des utilisateurs lambda dans un cadre incertain représente une nécessité pour que les personnes à la recherche d'information puissent obtenir des réponses fiables à leurs questions. Grâce aux représentations mathématiques offertes par les théories de l'incertain (la théorie des probabilités, des ensembles flous ou encore des fonctions de croyance), nous sommes capables de représenter et gérer les différents types d'imperfections reliées aux données.

C'est dans ce cadre que se situe notre travail, afin d'utiliser les apports proposés par ces théories pour classifier les différents types d'utilisateurs des sites communautaires et ainsi identifier les experts. Dans la suite de l'article, la section 2 présente les théories de l'incertain et plus précisément la théorie des ensembles flous et la théorie des fonctions de croyance employées dans ce travail. La section 3 s'intéresse à la détection des experts où nous présentons notre modélisation des utilisateurs, puis la comparaison de notre méthode de classification fondée sur la probabilité pignistique et celle fondée sur le K plus proche voisin crédibiliste sont présentées.

2. Les théories de l'incertain

Les théories de l'incertain sont issues de la précision des mathématiques classiques et de la subtile imprécision du monde réel. Plusieurs études ont permis d'aboutir à la théorie des ensembles flous ou des fonctions de croyance permettant de représenter l'imprécision et l'incertitude des connaissances.

La théorie des ensembles flous

La théorie des ensembles flous proposée d'abord par (Zadeh, 1965) avait pour objectif de sortir de la logique binaire en introduisant la notion d'une appartenance pondérée ou graduée. Autrement dit, permettre à un élément d'un sous-ensemble d'appartenir selon un degré plus ou moins fort à ce sous-ensemble en utilisant une fonction

appelée fonction d'appartenance. Soit un ensemble X, un sous-ensemble flou A de X est défini par : $\mu_A : X \longrightarrow [0, 1]$.

La théorie des fonctions de croyance

La théorie des fonctions de croyance initialement introduite par (Dempster, 1967), formalisée ensuite dans les travaux de (Shafer, 1976) est employée dans des applications de fusion d'information et de prise de décision. A partir d'un cadre de discernement Ω ($\Omega = \{\omega_1, \ldots, \omega_n\}$) qui est l'ensemble de toutes les hypothèses, nous définissons une fonction de masse sur l'ensemble de tous les sous-ensembles possibles de Ω à qui on affecte une valeur comprise entre $[0, 1]$ représentant ainsi sa masse de croyance élémentaire exprimée par $m : 2^\Omega \mapsto [0, 1]$. Formellement une fonction de masse m est définie par : $\sum_{X \subseteq \Omega} m(X) = 1$.

La règle de combinaison conjonctive proposée par (Smets, 1990) est utilisée lorsque les sources d'information sont considérées comme étant fiables ou lors de l'indépendance cognitive entre les données. Elle est donnée pour tout $X \in 2^\Omega$ par :

$$m_{conj}(X) = \sum_{Y_1 \cap Y_2 = X} m_1(Y_1) m_2(Y_2) \tag{1}$$

Pour la prise de décision, la transformation pignistique permet de transformer les fonctions de masse en mesures de probabilité. Ces fonctions permettent ainsi une prise de décision sur seulement des singletons. La probabilité pignistique définie par :

$$BetP(X) = \sum_{Y \in 2^\Omega, Y \neq \emptyset} \frac{|X \cap Y|}{|Y|} * \frac{m(Y)}{1 - m(\emptyset)}, \forall X \in 2^\Omega, X \neq \emptyset \tag{2}$$

3. Modélisation

Dans cette section nous décrivons le modèle proposé fonder sur l'utilisation des théories de l'incertain présentée dans la section 2. La modélisation des utilisateurs est fondée sur les fonctions d'appartenance permettant de représenter l'imprécision des informations. De ce fait, nous serons ainsi en présence d'un modèle imprécis offrant une vision imparfaite de la réalité. Ensuite, le passage aux probabilités, offre la modélisation de l'incertitude sur le modèle. Ces théories aussi riches soient elles sont des cas particuliers de la théorie des fonctions de croyance. Ainsi, cette dernière offre un cadre adéquat permettant de modéliser aussi bien, l'incertitude, l'imprécision que l'ignorance. C'est donc naturellement que nous avons choisi de construire notre travail sur ce formalisme qui offre un outil performant pour la fusion d'information. Nous considérons quatre types d'utilisateurs sur les sites de questions-réponses :

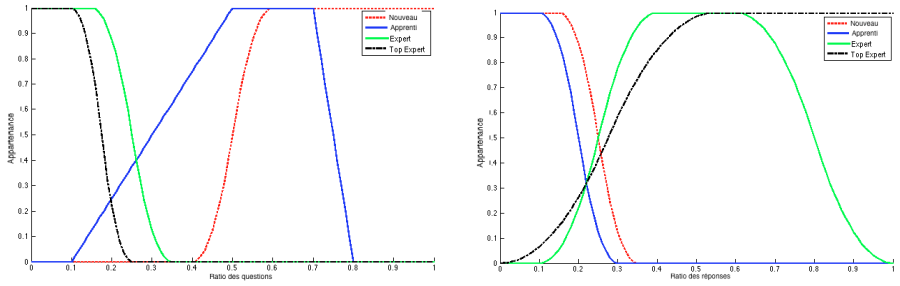
1. *Nouveau* : qui commence à apprendre les notions de base
2. *Apprenti* : qui a connaissances des notions de bases
3. *Expert* : donnant des réponses satisfaisantes et de bonne qualité
4. *Top Expert* : qui a de grandes connaissances et des réponses de bonne qualité

Nous reprenons ici l'hypothèse proposée par (Zhang *et al*., 2007) considérant le nombre de questions et de réponses comme étant des indicateurs de l'expertise d'une personne. Ils se fondent sur le fait que plus une personne pose des questions, plus elle manque de connaissances dans un domaine. Ainsi, dans ce cas, un utilisateur est considéré comme étant un *Nouveau* ou un *Apprenti*. Par ailleurs, plus un utilisateur fournit de réponses, plus il a d'expertise, et donc qu'il soit un *Expert* ou un *Top Expert*. Par conséquent, la modélisation de chaque utilisateur est directement reliée à deux variables relatives aux nombres de questions et de réponses. Ces variables sont mesurées à partir de deux rapports :

1. Rapport des questions : *Nombre de questions posées par utilisateur par thème*/ *Nombre total des questions posées par thème*
2. Rapport des réponses : *Nombre de réponses données par utilisateur par thème*/ *Nombre total des réponses données par thème*

Ici, nous proposons une modélisation probabiliste fondée sur les fonctions d'appartenance pour décrire le comportement des utilisateurs comme présenté dans la figure 1 selon les rapports des questions à droite et des réponses à gauche. Les fonctions d'appartenance permettent une modélisation précise du degré d'appartenance d'un utilisateur à une des classes.

A partir de ces fonctions d'appartenance, nous calculons les probabilités de chaque distribution décrivant le comportement des utilisateurs. Par la suite, nous construisons les fonctions de masse à partir du principe du moindre engagement. Ce dernier nous encourage à choisir la masse qui entraîne le moins de conséquences afin de ne pas présumer d'information que nous ne possédons pas. C'est une approche naturelle qui nous engage le moins possible vis à vis du choix des croyances. Les deux fonctions de masse ainsi obtenues une pour les questions l'autre pour les réponses sont combinées par la combinaison conjonctive de l'équation (1) de façon à renforcer les informations concordantes. Nous supposons ici l'indépendance cognitive entre le nombre de questions posées et le nombre de réponses données. La décision est ensuite prise par la probabilité pignistique offrant un compromis entre les différentes règles existantes.



(a) Représentation des Questions (b) Représentation des Réponses

FIGURE 1 – Fonctions d'appartenance

Nous avons modélisé les réponses et les questions en fonction du type d'utilisateur par les fonctions d'appartenance données dans la figure 1. Dans la figure 1a relative aux fonctions d'appartenance selon le rapport des questions, les deux premières courbes relatives aux *Experts* et aux *Top Experts*, nous avons fixé le degré d'appartenance pour un faible nombre de questions a une valeur maximale de $\mu = 1$, puis elle régressent progressivement jusqu'à avoir une valeur nulle. Ceci modélise qu'à partir d'un certain rapport entre les questions posées et le nombre total par thème, aucun utilisateur présumé appartenir à ces classes ne pose de question. Nous modélisons le comportement inverse pour les *Apprentis* et les *Nouveaux*. À partir d'un rapport de 0.7 un *Apprenti* ne pose plus de questions contrairement aux utilisateurs de la classe *Nouveau* qui pour un rapport de 1 ont une appartenance de 1. Nous conservons le même raisonnement pour les réponses mais avec une représentation différente comme décrit dans la figure 1b où l'appartenance d'un *Top Expert* a une valeur maximale de 1 et inversement pour les *Nouveaux* et les *Apprentis*.

4. Expérimentations

Nous avons généré : 999 *Nouveau*, 402 *Apprenti*, 101 *Expert*, 52 *Top Expert*, de façon à simuler un petit nombre d'experts et de top experts reflétant la réalité. Contrairement aux *Nouveaux* et aux *Apprentis* dont le nombre est relativement élevé car ils sont très nombreux sur les sites communautaires.

Nous procédons dans un premier temps à la classification des utilisateurs en employant la probabilité pignistique de l'équation (2). La matrice de confusion obtenue est présentée dans le tableau 1. Nous avons obtenus les probabilités de classification correcte avec une valeur de 0.8729 pour les *Nouveaux*, 0.5970 pour les *Apprentis*, 0.8020 pour les *Experts* et enfin 0.6154 pour les *Top Experts*.

Tableau 1 – Matrice de confusion avec BetP

	Nouveau	*Apprenti*	*Expert*	*Top Expert*
Nouveau	871	127	0	0
Apprenti	159	240	3	0
Expert	0	0	81	20
Top Expert	0	0	20	32

Afin d'évaluer notre méthode de classification, nous la confrontons au classificateur crédibiliste KNN de (Denoeux, 1995). La matrice de confusion est donnée dans le tableau 2. Nous avons obtenu les probabilités de classification correcte, avec une 0.8118 valeur pour les *Nouveaux*, 0.6517 pour les *Apprentis*, 0.7624 pour les *Experts* et enfin 0.4808 pour les *Top Experts*.

En comparant les résultats des tableaux 1 et 2 respectivement fondés sur la BetP et le E KNN, nous remarquons que notre méthode permet un meilleur taux de détection des utilisateurs de types *Experts* et *Top Experts* (0.8020 et 0.6154 avec le BetP contre

Tableau 2 – Matrice de confusion avec E KNN

	Nouveau	*Apprenti*	*Expert*	*Top Expert*
Nouveau	811	188	0	0
Apprenti	136	262	2	2
Expert	0	0	77	24
Top Expert	0	0	27	25

0.7624 et 0.4808 avec E KNN). Par ailleurs, nous remarquons que pour la détection des *Apprentis*, la deuxième méthode est légèrement meilleure avec une classification de 0.6517 et seulement 0.5970 pour la BetP. Ainsi, la classification fondée sur la BetP présente de meilleurs résultats pour la détection des *Experts* et des *Tops Experts* par rapport au K NN crédibiliste.

5. Conclusion

Dans cet article, nous avons proposé une représentation statistique fondée sur les théories de l'incertain des utilisateurs dans les réseaux communautaires. Nous avons combiné les informations relatives aux nombres de questions et réponses données pour chacun d'entre eux afin de permettre une prise de décision et une classification de chaque type d'utilisateur. Cette méthode fondée sur les probabilités pignistiques nous a permis d'avoir un meilleur taux de détection des experts que le E KNN. Comme perspectives à ce travail, nous allons non seulement explorer la notion des votes pour choisir les meilleures réponses mais aussi nous concentrer sur l'application de ce modèle sur des données réelles pour valider l'approche proposée.

Bibliographie

Bouguessa M., Dumoulin B., Wang S. (2008). Identifying authoritative actors in question-answering forums: the case of yahoo! answers. In, p. 866-874. ACM.

Dempster A. (1967). Upper and lower probabilities induced by a multivalued mapping. *Annals of Mathematical Statistics*, vol. 38, p. 325-339.

Denoeux T. (1995). A k-nearest neighbor classification rule based on Dempster-Shafer theory. *IEEE Transactions on Systems, Man and Cybernetics*, vol. 25, p. 804-813.

Gjergji K., Jurgen V G., D. S., Thore G. (2011). Cobayes: bayesian knowledge corroboration with assessors of unknown areas of expertise. In, vol. 2, p. 465-474.

Shafer G. (1976). *A mathematical theory of evidence*. Princeton University Press.

Smets P. (1990). The Combination of Evidence in the Transferable Belief Model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 12, n° 5, p. 447–458.

Zadeh L. A. (1965). Fuzzy sets. *Information and Control*, vol. 40, n° 8, p. 338–353.

Zhang J., Ackerman M. S., Adamic L. (2007). Expertise networks in online communities: structure and algorithms. In, p. 221–230. ACM Press.

Evidential community detection using structural and attribute information

Kuang Zhou [1,2], **Arnaud Martin** [2], **Quan Pan** [1]

1. *School of Automation, Northwestern Polytechnical University, Xi'an, Shaanxi 710072, PR China*
   kzhoumath@163.com
2. *DRUID, IRISA, University of Rennes 1, Rue E. Branly, 22300 Lannion, France*
   Arnaud.Martin@univ-rennes1.fr

RÉSUMÉ. *L'objectif de la détection de communautés est de créer une partition des sommets, de telle sorte que les communautés soient composées de sommets fortement connectés. Les approches existantes de détection de communautés se concentrent principalement sur la structure topologique du réseau, mais elles ignorent largement les informations disponibles à propos des attributs des nœuds. Dans cet article, une nouvelle appproche de détection de communautés qui utilise à la fois les informations structurelles et d'attributs pour extraire une structure de graphe imprécise, est proposée dans le cadre de la théorie des fonctions de croyance. L'objectif de notre méthode consiste à partitionner les sommets dans différents groupes afin que chaque cluster contienne un sous-graphe connecté densément avec les valeurs de l'attribut homogène. Les résultats expérimentaux montrent l'efficacité de la méthode proposée et montrent qu'elle pourrait améliorer les performances de la détection de communautés où les informations sur les propriétés du graphe sont disponibles en complément avec la structure topologique.*

ABSTRACT. *The goal of community detection is to partition nodes into different small subgroups in such a way that vertices in the same community have strong connections. Existing community detection approaches mainly focus on the topological structure of the network, but ignore the information about node attributes. In this paper, a new Evidential Community detection approach which could utilize both Structural and Attribute information, named ECSA, is proposed using belief functions to extract imprecise graph structure. The goal of our method is to partition vertices into different groups so that each cluster contains a densely connected subgraph with homogeneous attribute values. Experimental results illustrate the effectiveness of the proposed method and show that it could indeed improve the performance of community detection when the information about vertex properties is available together with topological structure.*

MOTS-CLÉS : *Détection de communautés; Attributs des nœuds; Communautés imprécises; Fonctions de croyance.*

KEYWORDS: *Community detection; Node attributes; Imprecise communities; Belief functions.*

DOI:10.3166/RIA. .1-?? © Lavoisier

Revue d'intelligence artificielle – n° / , 1-??

1. Introduction

Community detection is a useful unsupervised learning technique for detecting the cohesive groups in networks, and it has been developed rapidly in recent years and widely used in many applications. Traditional community detection approaches are mostly based on the topological structure of the graph. Sometimes, the vertex properties could also provide valuable information to guide the community detection process. For example, in a social network such as Facebook, users may have the following attributes: ID, a student/faculty status flag, gender, major, high school and college. Besides, traditional methods mostly focus on the non-overlapping communities. The work of detection overlapping communities incorporating both structural and attribute information has not been thoroughly studied yet. This is the motivation of our work.

In this paper, a new Evidential Community detection approach using both graph Structure and node Attributes (ECSA) is proposed in the framework of belief functions. The approach is based on an enhanced version of Evidential C-Means (ECM). An item is added into the objective function of ECM to reflect the consistence of members in the same group in terms of attribute information. The best partitions of the networks are obtained by optimizing the objective function. The experimental results demonstrate that the proposed approach could improve the performance of community detection when multiple information about the graph is available.

2. Related work

Detecting communities is still an open problem in social network analysis. Recently, significant progress has been achieved in this research field and several popular algorithms for community detection have been put forward. Among them we mention the modularity-based methods (Newman, 2006), label propagation algorithm (Raghavan *et al*., 2007), spectral optimization method (White, Smyth, 2005), and see (Fortunato, 2010) for review of the topic. However, these algorithms ignore the attributes of the nodes. Several new clustering methods that use both structure and attributes of graphs are introduced in recent years, such as SA-Cluster (Cheng *et al*., 2011), where a unified distance measure to combine structural and attribute similarities is defined, and then a clustering strategy similar to k-medoids is adopted to partition the nodes. Some other methods can be seen in the work of (Yang *et al*., 2013), (Ge *et al*., 2008), (Xu *et al*., 2012) and so on.

Recently, (Masson, Denoeux, 2008) proposed the application of evidential c-means (ECM) to get credal partitions for object data. The credal partition is a general extension of the crisp (hard) and fuzzy ones and it allows the object to belong to not only single clusters, but also any subsets of the set of clusters $\Omega = \{\omega_1, \cdots, \omega_c\}$ by allocating a mass of belief for each object in X over the power set 2^{Ω}. The additional flexibility brought by the power set provides more refined partitioning results than those by the other techniques allowing us to gain a deeper insight into the data.

3. Proposed method

Here we present the proposed ECSA algorithm in detail. ECSA is based on an enhanced version of ECM with Relational information (ECMwR). Therefore we will describe ECMwR first and then give the detailed ECSA algorithm.

3.1. ECMwR clustering

Let $\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n\}$ be a collection of vectors in $\mathbb{R}^p$ describing n objects to be classified into c clusters in the set $\Omega = \{\omega_1, \omega_2, \cdots, \omega_c\}$. Assume that some proximity information about the data set given in the form of relation matrix $\boldsymbol{W}_{n\times n}$ is available, and $w_{ij} \in \boldsymbol{W}$ represents the relationship between object $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$. Here we let $w_{ii} = 0, \forall i = 1, 2, \cdots, n$. The objective function of ECMwR is given as below:

$$\begin{aligned} J_{\text{ECMwR}} &= J_{\text{ECM}} + J_{\text{Re}} \\ &= \sum_{i=1}^{n} \sum_{A_h \subseteq \Omega, A_h \neq \emptyset} |A_h|^{\alpha}\, m_i^2(A_h) d_{ih}^2 + \sum_{i=1}^{n} \delta^2 m_i^2(\emptyset) + \tau \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\mathcal{K}_{ij} \end{aligned} \tag{1}$$

with

$$\mathcal{K}_{ij} = \begin{cases} \sum\limits_{A_e \cap A_r = \emptyset} m_i\left(A_e\right) m_j\left(A_r\right) & i \neq j, \\ 0 & i = j. \end{cases} \tag{2}$$

The term $\mathcal{K}_{ij}$, which in fact is the mass assigned to the empty set by the conjunctive combination, reflects the disagreement between masses for $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$. Parameter α is a tuning parameter allowing to control the degree of penalization for subsets with high cardinality, and d_{ih} denotes the distance (generally Euclidean distance) between x_i and the barycenter (*i.e.* prototype) associated with A_h, and τ balances the contribution of two components, *i.e.,* J_{ECM} and J_{Re}. Parameter δ is used to detect outliers. The objective function of J_{ECMwR} should be subject to constraints in Eq. (3).

$$\sum_{A_j \subseteq \Omega, A_j \neq \emptyset} m_i(A_j) + m_i(\emptyset) = 1, m_i\left(A_j\right) \geq 0, \;\; m_i\left(\emptyset\right) \geq 0. \tag{3}$$

To solve the constrained minimization problem, the method of Lagrange multipliers provides a classical way. An alternate optimization scheme similar to that in FCM and ECM algorithms can be designed for ECMwR with this method. First, we consider that the prototype set of clusters, V, is fixed. The update equations of $m_{ij} \triangleq m_i(A_j)$ for ECMwR could be derived as follows.

$$m_{ik} = m_{ik}^{\text{ECM}} + \tau m_{ik}^{\text{Re}} \tag{4}$$

with

$$m_{ik}^{\text{ECM}} = \frac{|A_k|^{-\alpha} d_{ik}^{-2}}{\sum\limits_{A_h \subseteq \Omega, A_h \neq \emptyset} |A_h|^{-\alpha} d_{ih}^{-2} + \delta^{-2}}, \;\; \forall i = 1, 2 \cdots, n, \forall k / A_k \subseteq \Omega, A_k \neq \emptyset, \tag{5}$$

$$m_{ik}^{\text{Re}} = \frac{\dfrac{\sum\limits_{A_h \subseteq \Omega, A_h \neq \emptyset} \sum_{j=1}^{n} w_{ij} \sum\limits_{A_h \cap A_l = \emptyset} m_{jl} |A_h|^{-\alpha} d_{ih}^{-2} + \sum\limits_{j=1}^{n} w_{ij}\delta^{-2}}{\sum\limits_{A_h \subseteq \Omega, A_h \neq \emptyset} |A_h|^{-\alpha} d_{ih}^{-2} + \delta^{-2}} - \sum_{j=1}^{n} w_{ij} \sum\limits_{A_k \cap A_l = \emptyset} m_{jl}}{|A_k|^{\alpha} d_{ik}^2}, \tag{6}$$

and

$$m_{i\emptyset} = 1 - \sum_{A_j \neq \emptyset} m_{ij}, \ \ \forall i = 1, 2, \cdots, n. \tag{7}$$

It is remarkable that Eq. (4) is a group of equations of m_{ik}, and the constraints are not explicitly satisfied. To obtain the solution of Eq. (4), the simple successive-substitution method, in which one can repeatedly use old values of m_{ik} in Eq. (6) to get m_{ik}^{Re} and then solve for new values of m_{ik} from Eq. (4) until convergence, could be utilized. In practice, one can improve the order of convergence of this approach by the application of Seidel iteration scheme, where all the new available mass values are used for solving m_{ik}.

As we can see, the update formula of m_{ik} derived in Eq. (4) does not guarantee non-negativity. The Karush–Kuhn–Tucker (KKT) conditions could be used to force the memberships to be positive. Since the application of KKT conditions would yield more complicated update equations, in this paper we use a simple clipping strategy: at each iteration set those negative mass values obtained by Eq. (4) to 0, and renormalize the values to sum to 1.

From Eq. (1) it is easy to know the penalty term J_{Re} in the objective function of ECMwR does not depend on the cluster centroids. Thus, given the partition matrix $\boldsymbol{M}$, the updating of the prototype set $\boldsymbol{V}$ in ECMwR could be evoked by the same scheme as that in the application of ECM (Masson, Denoeux, 2008).

3.2. Community detection approach

An attributed graph is denoted as $G = (V, E, A)$, where V is the set of nodes, E is the set of edges, and $A = \{a^1, a^2, \cdots, a^d\}$ is the set of d attributes associated with nodes in V. Each vertex v_i is associated with an attribute vector $\left(a_i^1, \cdots, a_i^d\right)$. ECSA algorithm is described in the following.

(1) Map the topological structure to feature vectors by some spectral methods;

(2) Construct the relational matrix based on the attribute values; In this work we only consider discrete attributes. The similarity between two nodes could be determined by examining each of d attributes and counting the number of attribute values they share in common;

(3) Evoke ECMwR clustering algorithm and obtain the detected communities.

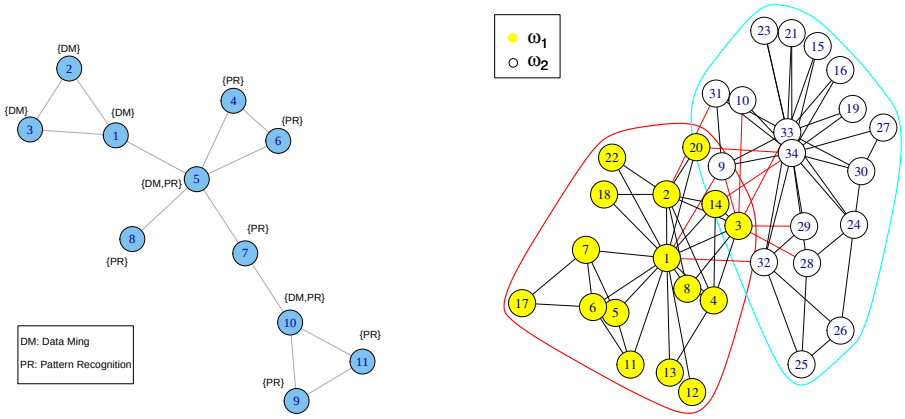
Remark: As in ECM, in ECMwR the number of parameters to be optimized is exponential and depends on the number of clusters. For the number of classes larger than 10, calculations are not tractable. But we can consider only a subclass with a limited number of focal sets. For instance, we could constrain the focal sets to be composed of at most two specific classes.

4. Experiments

To show the principle of the proposed method, a small illustrative example of a co-authorship network displayed in Figure 1.a is first considered. In the graph each vertex represents an author while an edge represents the co-author relationship between two

authors. In addition, there are primary topics associated with each author. The research topic is regarded as an attribute describing the vertex property. The graph structure is mapped into Euclidian space using signal prorogation method (Hu *et al.*, 2008) and the vertex attributes are used to construct the relation matrix w_{ij}. For the author i and j, if they share r same topics, $w_{ij} = r$; Otherwise $w_{ij} = 0$.

Two other community detection schemes are compared. One is FCM based clustering after the spectral mapping, the other is ECM based clustering. Note that in these two algorithms, only the graph structure is used. The authors are assigned to the groups with maximal mass value by ECM and ECMwR. The results show that by FCM and ECM the authors in the same detected group by FCM and ECM may have different topics. On the contrary, authors in the same found community by ECMwR are not only closely connected, but also sharing homogeneous research topics. Credal partitions enable us to detect the imprecise overlapped nodes. The overlapped node found by ECM is node 7, while node 5 is regarded as overlapping by ECMwR. Author 5 and Author 10 regard both DM (Data Mining) and PR (Pattern recognition) as their own topics, but only Author 5 is clustered into the imprecise cluster by ECMwR. This is due to the fact that Author 5 has collaborated with authors in both two communities while Author 10 only has co-authorship with those whose topic is PR. The detection result by ECMwR is more reasonable as it takes advantage of the structural and attribute similarities at the same time.



a. Co-Authorship Network b. Karate Club Network

Figure 1. Original networks.

In the next experiment we will use a widely used benchmark in detecting community structures, “Karate Club", studied by Wayne Zachary. The network consists of 34 nodes and 78 edges representing the friendship among the members of the club (see Figure 1.b). In the original network there is no attribute for nodes. We generate random attributes (a_{i1}, a_{i2}) for node i based on the ground-truth. For all the nodes in community ω_1, $a_{i1} = 1, a_{i2} = 0$, while for those in ω_2, $a_{i1} = 0, a_{i2} = 1$. Then we modify the attribute of node 3 to $(1, 1)$. From the results we see, without the attribute information, nodes 3, 9, 14 are all partitioned into the imprecise class $\{\omega_1, \omega_2\}$ by

ECM. But after taking the attribute information into account, only node 3 is regarded as a member in $\{\omega_1, \omega_2\}$. This is due to the fact that node 3 lies in the overlap in terms of not only topological structure, but also attribute values.

From the two experiments, we can get that: 1. The found communities by ECSA contains members not only being connected closely but also sharing similar attributes. 2. ECSA could detect overlapping communities in the concept of credal partitions. Also for the members in the overlap, they are not only frequently connected to the vertices in all the related groups but also labelled with more than one attribute.

5. Conclusion

In this study, an evidential community detection method incorporating both structural and attribute information is presented in the framework of belief functions. The proposed algorithm is based on an enhanced version of ECM clustering using available relational information. Experimental results show that our method will provide detected communities with nodes not only being densely connected but also have homogeneous attribute values.

Bibliographie

Cheng H., Zhou Y., Yu J. X. (2011). Clustering large attributed graphs: A balance between structural and attribute similarities. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 5, n° 2, p. 12.

Fortunato S. (2010). Community detection in graphs. *Physics Reports*, vol. 486, n° 3, p. 75–174.

Ge R., Ester M., Gao B. J., Hu Z., Bhattacharya B., Ben-Moshe B. (2008). Joint cluster analysis of attribute data and relationship data: The connected k–center problem, algorithms and applications. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 2, n° 2, p. 7.

Hu Y., Li M., Zhang P., Fan Y., Di Z. (2008). Community detection by signaling on complex networks. *Physical Review E*, vol. 78, n° 1, p. 016115.

Masson M.-H., Denoeux T. (2008). Ecm: An evidential version of the fuzzy c-means algorithm. *Pattern Recognition*, vol. 41, n° 4, p. 1384–1397.

Newman M. E. (2006). Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, vol. 103, n° 23, p. 8577–8582.

Raghavan U. N., Albert R., Kumara S. (2007). Near linear time algorithm to detect community structures in large-scale networks. *Physical Review E*, vol. 76, n° 3, p. 036106.

White S., Smyth P. (2005). A spectral clustering approach to finding communities in graph. In *Sdm*, vol. 5, p. 76–84.

Xu Z., Ke Y., Wang Y., Cheng H., Cheng J. (2012). A model-based approach to attributed graph clustering. In *Proceedings of the 2012 ACM SIGMOD international conference on management of data*, p. 505–516.

Yang J., McAuley J., Leskovec J. (2013). Community detection in networks with node attributes. In *Data mining (ICDM), 2013 ieee 13th international conference on*, p. 1151–1156.

Ensemble selection for community detection

Parisa Rastin [1], **Rushed Kanawati** [2]

LIPN CNRS UMR 7030
Sorbonne Paris Cité, Université Paris 13
Villetaneuse, France
surname.name@lipn.fr

RÉSUMÉ.

ABSTRACT. Ensemble clustering approaches have been recently applied, in a variety of ways, in order to enhance the quality and/or the execution time of community detection approaches. The quality gain that can be obtained from applying ensemble approaches is known to be tightly linked to both quality and diversity of the input base clusterings. Clustering ensemble selection approaches are devised in order to select the most promising base clustering to combine. However, In complex network analysis context, most of existing work simply ignore this important issue of ensemble selection. In this paper we intend to fill this gap. We propose a graph-based ensemble selection approach that allows to take into account an ensemble *of quality and diversity criteria. First experiments on benchmark networks show the validity of our approach.*

MOTS-CLÉS : Sélection d'ensemble, réseaux multiplexes, détection de communuatés

KEYWORDS: Ensemble clustering, Clustering ensemble selection, Multiplex network, Community detection.

DOI:10.3166/RIA..1-6

1. Introduction

Community detection is a central task in the field of complex network analysis and mining. A community is often defined as a dense subgraph that is loosely linked to other communities in the network. Unfolding the community structure of a network is a first step towards understanding the complex interaction patterns that can be observed in real world applications. A wide variety of community detection approaches have been proposed in the literature (Fortunato, 2010 ; Tang, Liu, 2010). Different algorithms have different execution times and yield results of various quality. Generally low time complexity algorithms show also low robustness. This is for instance the case of the *Louvain* approach (Blondel *et al*., 2008) and the high speed label propagation algorithm (Raghavan *et al*., 2007). Ensemble clustering (EC) approaches are then proposed as mean to cope with the robustness issue. One can compute different unstable clusterings, using a fast but unrobust algorithm, and then combine them in a more robust and accurate consensus clustering (Asur *et al*., 2007 ; Seifi, Guillaume, 2012 ; Lancichinetti, Fortunato, 2012). EC approaches have been applied to different tasks including: computing communities cores (Seifi, Guillaume, 2012), computing dynamic communities (Lancichinetti, Fortunato, 2012), multi-objective local communities identification (Kanawati, 2015a), community detection in multiplex networks (Hmimida, Kanawati, 2015), and large-scale graph coarsening (Staudt, Meyerhenke, 2013 ; Kanawati, 2015b).

Different consensus clustering functions have been proposed in the literature. Existing functions can be roughly classified into two classes: *evidence accumulation based functions* (Fred, Jain, 2005) and *graph-based functions* (Strehl, Ghosh, 2003). One widely applied graph-based approach is the *CSPA* algorithm which is based on constructing a **consensus graph** out of the set of partitions to be combined (Fern, Brodley, 2004). The consensus graph G_{cons} is defined over the same set of clustered data items. Two nodes $v_i, v_j \in V$ are linked in G_{cons} if there is at least one base partition where both items i,j are in a same cluster. Each link (v_i, v_j) is weighted by the frequency of instances that nodes v_i, v_j are placed in the same cluster. Links in the obtained consensus graph whose weights (frequency) are under a given threshold $\alpha \in [0, 1]$ are pruned yielding decomposing the graph in a set f connected components. These connected components represent the consensus clustering.

Recently, different works have showed that the quality of the output of an EC approach is tightly related to both the *quality* of each partition in the base clustering set and *diversity* of these clusterings. Cluster ensemble selection (CES) approaches have been proposed in order to compute a subset of the base clusterings set that maximize both the quality and the diversity (Akbari *et al*., 2015 ; Azimi, Fern, 2009 ; Fern, Lin, 2008)

The *diversity* of clusterings can be estimated by applying different external cluster evaluation indexes or cluster dissimilarity indexes such as: rand index and the adjusted rand index (ARI) (Hubert, Arabie, 1985), Normalized mutual information (NMI) and Information variation indexes (Meila, 2003). Specific versions of these indexes

can be used for clustering in networks (Labatut, 2012). The quality of a clustering can be evaluated by different *internal* evaluation indexes. Examples are the modularity (Newman, M. Girvan, 2004), or the different local modularities functions (Yang, Leskovec, 2012 ; Kanawati, 2015a).

Almost all CES approaches require the number of base clusterings to select as an input (Azimi, Fern, 2009). Some are based on selecting high quality base clusterings. Some compute a trade-off between quality and diversity (Fern, Lin, 2008). However, all existing approaches apply one index for measuring the quality and another for evaluating the diversity of clusterings. In this work we investigate the use of an *ensemble* of quality and diversity indexes for CES. In addition the devised approach computes automatically the number of base clusterings to be selected. This will be explained in more details in next section.

2. Multiplex network based CES

Algorithm 1 sketches the outlines of the proposed approach. The basic idea is to define a multiplex network over the provided set of base clusterings. A multiplex network is multi-slice network where each slice contains the same set of nodes but different kinds of links (Hmimida, Kanawati, 2015). In our case, each slice of the defined multiplex is modeled by a *proximity graph* constructed by applying a given clustering dissimilarity index (ex. NMI, ARI, VI). Different types of proximity graphs can be used. In this work, we first explore using *relative neighborhood graphs* (RNG) (Toussaint, 1980). Though the complexity of RNG graph constriction is relatively high, the resulted graph is proved to be connected and sparse. Recall also that the graph is defined over the set of base clusterings which cardinality is usually low.

Algorithm 1 Graph-based cluster ensemble selection algorithm

Require: $\Pi = \{\pi_1, \ldots, \pi_r\}$ a set of base clusterings
Require: $\mathcal{S} = \{S_1, \ldots, S_n\}$ A set of partition similarity functions
Require: $\mathcal{Q} = \{Q_1, \ldots, Q_m\}$ A set of partition quality functions
1: $\Pi^* \leftarrow \emptyset$
2: $MUX \leftarrow$ **Multiplex**(Π)
3: **for all** $S_i \in \mathcal{S}$ **do**
4: $\quad MUX$.**add_layer**(proximity_graph(Π, S_i))
5: **end for**
6: $\mathcal{C} = \{c_1, \ldots, c_k\} \leftarrow$ **community_detection**(MUX)
7: **for all** $c \in \mathcal{C}$ **do**
8: $\quad \hat{\pi} \leftarrow$ **ensemble_Ranking**(c, $\mathcal{Q}$)
9: $\quad \Pi^* \leftarrow \Pi^* \cup \{\hat{\pi}\}$
10: **end for**
11: **return** Π^*

A community detection algorithm is applied to the obtained multiplex network. Recall that a community is defined as a dense sub-graph that is loosely connected

to other communities in the network. Different approaches for community detection in multiplex networks can be applied. A survey on such algorithms is provided in (Kanawati, 2015c). In this work we apply a seed-centric approach proposed in (Hmimida, Kanawati, 2015). Since two nodes (clusterings) are linked if they very similar, a community in the multiplex network delimits a number of base clusterings that are similar among them and diverse in regard to other clusterings belonging to other communities. We can then stress the diversity of clusterings to return by selecting one clustering from each detected community. We rank clusterings in each community applying an ensemble-ranking approach using different (internal) clustering quality indexes. From each community we select the base clustering ranked at the top. First experiments on benchmark datasets shows the effectiveness of the proposed CES approach. This leads to select high quality but diverse base clusterings.

3. Experiments

As a first evaluation of the proposed CES approach, we have conducted the following primary experiment. We have selected a set of benchmark networks frequently used in works dealing community detection in complex networks and for which we have a ground-truth decomposition into communities. These networks are the following: the Zachary Karate club network, the US politics books network and Dolphins network (Kanawati, 2015a). To each network, we apply the *label propagation* community detection algorithm 100 times (Raghavan *et al*., 2007). This algorithm is known to be quick but highly instable. We then obtain a set of 100 different clusterings that compose our raw base clusterings set. We then compared the results of applying a CSPA ensemble clustering approaches directly to the raw base clusterings set to those obtained by applying the same ensemble clustering algorithm to the subset obtained after applying our CES approach. For the CES algorithm, we used the modularity, and the conductance as a clustering quality indexes. NMI, ARI and VI are used to measure clustering dissimilarity (diversity). The *muxLicod* algorithm (Hmimida, Kanawati, 2015) is applied in order to compute communities in the obtained multiplex network. A simple Borda rank aggregation method is applied in order to select the top quality clustering from each detected community. The results are evaluated in function similarity of obtained clustering to the ground-truth clustering using again the NMI and ARI indexes. The modularity (Q) is also used to evaluate the overall quality of obtained results. As shown in next table, for all three networks, the CES approach does enhance the obtained results. These first results are encouraging. But the work is still in its early stages. Experiments on larger datasets and using different quality and diversity indexes are scheduled. The effect of the choice of the multiplex community detection algorithm should also be studied. Another factor to analyse is the enhancement in using an ensemble of indexes rather than using single quality/diversity index should also be done.

Tableau 1. Evaluation of the proposed graph-based ensemble selection

Dataset	Approach	NMI	ARI	Q	# Communities
Zachary	EC	0.57	0.46	0.40	5
	CES + EC	**0.77**	0.69	0.34	2
US Politics	EC	0.55	0.68	0.51	5
	CES+EC	**0.68**	0.67	0.42	6
Dolphins	EC	0.55	0.39	0.51	5
	CES +EC	**0.58**	0.59	0.53	3

4. Conclusion

In this work, we have proposed a new approach for enhancing the output of ensemble clustering by applying an original ensemble selection process. The approach consists in applying a community detection algorithm to a multiplex graph defined over the set of base clustering to filter. First results show that the overall quality of obtained clustering is enhanced when applying ensemble selection process. Experiments on large-scale datasets are planned in order to confirm these first but promising results. Comparisons with other ensemble selection approaches based on implicit quality estimation are also scheduled.

Bibliographie

Akbari E., Dahlan H. M., Ibrahim R., Alizadeh H. (2015). Hierarchical cluster ensemble selection. *Engineering Applications of Artificial Intelligence*, vol. 39, p. 146-156.

Asur S., Ucar D., Parthasarathy S. (2007). An ensemble framework for clustering protein-protein interaction networks. In *Ismb/eccb (supplement of bioinformatics)*, p. 29-40.

Azimi J., Fern X. (2009). Adaptive cluster ensemble selection. In C. Boutilier (Ed.), *Ijcai*, p. 992-997.

Blondel V. D., Guillaume J.-l., Lefebvre E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2008, p. P10008.

Fern X. Z., Brodley C. E. (2004). Solving cluster ensemble problems by bipartite graph partitioning. In C. E. Brodley (Ed.), *Icml*, vol. 69. ACM.

Fern X. Z., Lin W. (2008). Cluster ensemble selection. *Statistical Analysis and Data Mining*, vol. 1, n° 3, p. 128-141.

Fortunato S. (2010). Community detection in graphs. *Physics Reports*, vol. 486, n° 3-5, p. 75-174.

Fred A. L. N., Jain A. K. (2005). Combining multiple clusterings using evidence accumulation. *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, n° 6, p. 835-850.

Hmimida M., Kanawati R. (2015, March). Community detection in multiplex networks: A seed-centric approach. *Networks and Heterogeneous Media*, vol. 10, n° 1, p. 71-85. (Special Issue on New trends, models and applications in Complex and Multiplex Networks)

Hubert L., Arabie P. (1985). Comparing partitions. *Journal of classification*, vol. 2, n° 1, p. 192-218.

Kanawati R. (2015a, February). Empirical evaluation of applying ensemble methods to ego-centered community identification in complex networks. *Neurocomputing*, vol. 150, B, p. 417-427.

Kanawati R. (2015b, April). Ensemble selection for enhancing graph coarsening quality. In *Proceedings of 5th international workshop on social network analysis.* Capri.

Kanawati R. (2015c, July). Multiplex network mining (a tutorial). In *5th international conference on web intelligence, mining and semantics.* Larnaca.

Labatut V. (2012). Une nouvelle mesure pour l'évaluation des méthodes de détection de communauté. In *Actes de 3ième conférence sur les modèle et analyse des réseaux: approches mathématiques et informatiques (marami'12).*

Lancichinetti A., Fortunato S. (2012). Consensus clustering in complex networks. *Sci. Rep.*, vol. 2.

Meila M. (2003). Comparing clusterings by the variation of information. In B. Schölkopf, M. K. Warmuth (Eds.), *Colt*, vol. 2777, p. 173-187. Springer.

Newman M. J., M. Girvan M. (2004). Finding and evaluating community structure in networks. *Physics review E*, vol. 69, p. 026113:1–022613:15.

Raghavan U. N., Albert R., Kumara S. (2007, September). Near linear time algorithm to detect community structures in large-scale networks. *Physical Review E*, vol. 76, p. 1-12.

Seifi M., Guillaume J.-L. (2012). Community cores in evolving networks. In A. Mille, F. L. Gandon, J. Misselis, M. Rabinovich, S. Staab (Eds.), *Www (companion volume)*, p. 1173-1180. ACM.

Staudt C., Meyerhenke H. (2013). Engineering high-performance community detection heuristics for massive graphs. In *Icpp*, p. 180-189. IEEE.

Strehl A., Ghosh J. (2003). Cluster ensembles: a knowledge reuse framework for combining multiple partitions. *The Journal of Machine Learning Research*, vol. 3, p. 583-617.

Tang L., Liu H. (2010). *Community detection and mining in social media*. Morgan & Claypool Publishers.

Toussaint G. T. (1980). The relative neighbourhood graph of a finite planar set. *Pattern Recognition*, vol. 12, n° 4, p. 261-268.

Yang J., Leskovec J. (2012). Defining and evaluating network communities based on ground-truth. In M. J. Zaki, A. Siebes, J. X. Yu, B. Goethals, G. I. Webb, X. Wu (Eds.), *Icdm*, p. 745-754. IEEE Computer Society.

Apprentissage collaboratif de proximité

Lise-Marie Veillon [1], **Henry Soldano** [1,3], **Gauvain Bourgne** [2]

1. *Université Paris 13, Sorbonne Paris Cite, L.I.P.N UMR-CNRS 7030 F-93430, Villetaneuse, France*
   veillon@lipn.univ-paris13.fr,henry.soldano@lipn.univ-paris13.fr
2. *Sorbonne Universités, UPMC Univ Paris 06, CNRS, LIP6 UMR 7606, 4 place Jussieu 75005 Paris,France*
   gauvain.bourgne@lip6.fr
3. *Atelier de BioInformatique, UPMC, F-75005, Paris, France*

RÉSUMÉ. *Nous nous intéressons à l'apprentissage collaboratif dans une société d'agents organisés en réseau. Plus précisément, chaque agent révise son modèle courant à partir de nouvelles observations (exemples) rendant celui-ci incohérent. Lors d'une révision chaque agent peut bénéficier, à travers des interactions, des observations de ses voisins et mémoriser celles-ci. Du fait qu'un agent ne communique directement qu'avec ses voisins, la vitesse d'apprentissage dépendra des observations disponibles dans son voisinage et des caractéristiques du réseau liant les agents : en effet la structure du réseau influe sur la circulation des observations dans celui-ci lors de ces interactions. La vitesse d'apprentissage dépend également de l'attitude de l'agent vis à vis des modèles qu'il critique : nous considérons ici principalement le cas où un modèle qui n'est plus critiquable par ses voisins est adopté par ceux-ci, mais aussi celui où un agent garde toujours pour modèle courant le modèle qu'il a lui-même révisé en dernier lieu. L'idée générale est de comprendre quelles sont les caractéristiques du réseau qui influent le plus sur le temps nécessaire à la collectivité pour diffuser des informations et construire des modèles cohérents avec ceux-ci.*

ABSTRACT.

MOTS-CLÉS : *Intelligence Artificielle, Apprentissage, Collectif, Symbolique, Graphes, Réseaux, Multi-Agents, Communication Limitée*

KEYWORDS: *Artificial intelligence, Learning, Collective, Symbolic, Graphs, Networks, Multi-Agents, Limited communication*

DOI:10.3166/RIA.??.1-?? © 2015 Lavoisier

Revue d'intelligence artificielle – n° ?/2015, 1-??

1. Introduction

Les réseaux sociaux constituent de très grands réseaux recueillant régulièrement de nouvelles informations en tout point du réseau. Le protocole SMILE (Sound Multi-agents Incremental LEarning)(Bourgne, Bouthinon *et al*., 2009) est un protocole d'apprentissage collaboratif dans un système multi-agents (SMA) c'est-à-dire une tentative de modélisation par simulation, proche de celle de Ontañón (Ontañón, Plaza, 2010), des phénomènes d'apprentissage collaboratif par échange d'observations (faits) et transmission d'hypothèses (théories) dans un réseau social. Dans cet article nous nous intéressons aux caractéristiques principales ou secondaires du réseau influant sur la vitesse d'apprentissage d'une théorie (ici une formule-cible). Plus généralement ce travail se situe dans le cadre de l'étude des phénomènes d'apprentissage dans les réseaux sociaux. Il y a relativement peu de travaux sur ce sujet dont l'épistémologie de réseau de Zollman (Kevin J.S. Zollman, 2007) qui modélise la propagation et la sélection d'hypothèses dans un société plutôt que leur formation. Par ailleurs la propagation d'informations dans les réseaux a été largement étudiée (Guille *et al*., 2013) mais pas le lien avec les phénomènes de formation de connaissances.

Dans SMILE les agents possèdent un mécanisme de révision d'une hypothèse couvrant les exemples positifs et rejetant les exemples négatifs d'un concept cible. Chaque agent est autonome et dispose de sa propre mémoire d'exemples et son hypothèse courante. Au fur et à mesure de l'arrivée des exemples les agents doivent maintenir la cohérence de leur hypothèse avec les exemples qu'ils ont reçus personnellement d'une part et ceux de l'ensemble de la communauté d'autre part. Lorsqu'un agent voit son hypothèse actuelle contredite par l'arrivée d'un nouvel exemple, il la révise puis la confronte à la critique des autres agents. Les réponses sont soit un contre exemple qui va provoquer de nouveau une révision soit une acceptation si l'hypothèse est en accord avec tous les exemples en mémoire. Quand les réponses sont toutes positives, l'agent apprenant partage son hypothèse une dernière fois en indiquant qu'elle est communément acceptée. A cette occasion tous les agents adoptent ou non cette hypothèse, qui devient alors commune. D'autre part, l'apprenant choisit d'oublier ou non les exemples qu'il a reçu de ses voisins. Dans une version ultérieure de SMILE (Bourgne, El Fallah Seghrouchni, Soldano, 2009) la topologie du réseau n'est pas contrainte. Un mécanisme de propagation des hypothèses à l'ensemble du réseau permet de préserver la cohérence globale du système. Cependant, ce mécanisme est coûteux et dans la présente étude nous évitons d'imposer cette propagation. Nous donnons ci dessous un exemple de révision d'une hypothèse, sans propagation, par un agent SMILE:

EXEMPLE 1. —

On se place dans un cas de quatre agents alignés où l'objectif est d'apprendre la formule-cible A ou B à partir d'exemples positifs (satisfaisant la formule cible) et négatifs (ne satisfaisant pas la formule cible). L'hypothèse courante d'un agent doit donc couvrir les exemples positifs (e^+) et rejeter les exemples négatifs (e^-) connus par l'agent. Elle est alors considérée comme cohérente avec ces exemples. Après que l'agent 1 et l'agent 3 aient reçu respectivement les exemples positifs $e_1^+(A \wedge \neg B \wedge C)$

et $e_2^+(A \wedge B \wedge \neg C)$ puis révisé leurs hypothèses en conséquence, l'agent 2 reçoit un exemple positif $e_3^+(\neg A \wedge B \wedge C)$ (figure 1). Il ajoute e_3^+ à sa mémoire d'exemple et constate une contradiction avec son hypothèse courante. A la suite de sa révision il forme l'hypothèse H3 qu'il soumet à la critique de son voisin Agent 1. Celui ci dispose d'un contre exemple e_1^+ non couvert par H3 et l'envoie comme réponse à l'agent 2. A l'issu d'une deuxième révision tenant compte de e_1^+ ajouté en mémoire, Agent 2 propose sa nouvelle hypothèse H4 successivement à Agent 1 et Agent 3 ses voisins qui l'acceptent faute de contre exemple à renvoyer. L'hypothèse H4 étant cohérente avec la mémoire de Agent 2 et ses voisins, il a terminé sa révision globale. Le protocole utilisé comprends l'adoption d'hypothèse mais pas d'oubli. Agent 1 et Agent 3 étant voisins de Agent 2 ils adoptent sa nouvelle hypothèse H4. Agent 2 n'étant pas soumis à oubli, il conserve son exemple externe e_1^+.

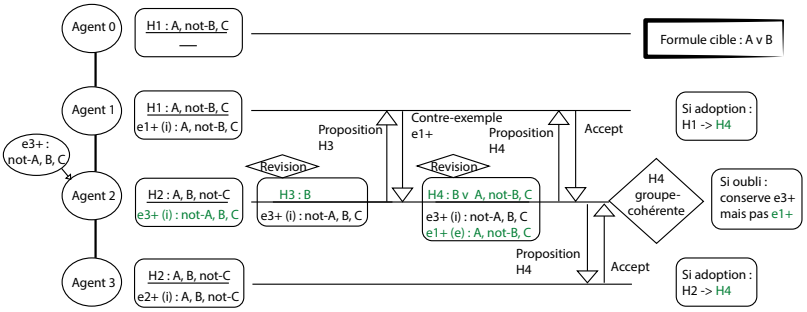


Figure 1. Exemple d'apprentissage collectif local dans un réseau de 4 agents alignés avec adoption et sans oubli. (i) : exemple interne ; (e) : exemple externe

□

Dans cet article, nous étudions comment la précision de l'apprentissage (probabilité que le modèle soit correcte relativement à une nouvelle observation) varie en fonction des caractéristiques du réseau. Plus précisément, nous examinons la précision de l'apprentissage en fonction du nombre d'observations perçues directement par les agents du SMA. De même, nous examinons les mécanismes de mémorisation des exemples et d'adoption des hypothèses favorisant la précision de l'apprentissage lorsque plusieurs agents révisent simultanément.

2. Conditions expérimentales

Afin de mettre en avant les particularités de réseaux qui favorisent plus ou moins l'apprentissage collectif, un certain nombre d'entre eux ont été étudiés. Chaque réseau constitue un graphe de communication dont les agents sont les sommets et les arêtes qui les relient indiquent la possibilité qu'ils ont de communiquer. Les graphes sont fixes dans le temps et non orientés. Cette étude se limite à un nombre constant d'agents de 50 individus. Le problème d'apprentissage choisi, le multiplexer11, est une formule booléenne difficile par son morcellement de l'espace des hypothèses possibles et pour laquelle la collectivité se montre particulièrement efficace. L'apprentissage est évalué

ici par la justesse moyenne des agents , c'est-à-dire la proportion moyenne d'exemples bien classés dans un ensemble test par les agents. La justesse moyenne donnée résulte de 100 expériences, avec des ensembles d'exemples non bruités différents. Au cours des expériences des graphes variés sont utilisés (Veillon, 2015). Parmi ceux-ci un certain nombre d'arbres réguliers à k fils par nœud nommés Arb'k, un graphe régulier circulaire Reg'n' construit en reliant chaque sommet aux n/2 précédents, un donut , en forme de tore de longueur 10 pour une section de 5 agents en cercle. Les smallWorlds SmW'd'p'x' sont générés à partir de graphes réguliers circulaires de degré moyen d pour lesquels toutes les arêtes ont une probabilité x d'avoir été modifiée à l'une de ses extrémités, les multipôles sont constitués de sommets prioritaires, les pôles, et de sommets secondaires. Ces graphes sont construits en reliant les pôles à un plus grand nombre d'autres sommets possibles. Les multipôles relient en premier les pôles entre eux tandis qu'un multipôle dit séparé met la priorité sur leurs liaisons avec les sommets secondaires. Enfin les MCluster5'circ/star' sont constitués de dix cliques de 5 sommets dont une arête a pu être déplacée pour rendre le graphe connexe en forme de cercle ou d'étoile.

3. Exploration de la vitesse d'apprentissage en fonction des caractéristiques des graphes

En premier lieu des graphes similaires de type smallWorlds sont sélectionnés pour être comparés. Le premier paramètres permettant la génération de ces graphes est le degré moyen. Étant donné le nombre fixe d'agents dans cette étude la variation du degré moyen est assimilée à celle de la densité du graphe ($densite = degreMoyen/49$). La figure 2 (gauche) montre que la vitesse d'apprentissage augmente avec la densité du graphe. La densité peut se comprendre comme la proportion des communications présentes dans ce réseau par rapport au cas complet de la clique. Plus les limitations sont importantes, moins l'apprentissage est efficace. Ce n'est cependant pas le seul critère puisque SmW4p01 et SmW4p05 ont des résultats différents pour une même densité. La reconnexion aléatoire d'un plus grand nombre d'arêtes pour SmW4p05 a notamment l'effet de diminuer le diamètre et la distance moyenne entre deux sommets.

Afin de confirmer cette tendance, nous comparons un échantillon varié de graphes de degré moyen 4. (figure 2 (droite)) Les structures privilégiées sont de nouveau celles présentant des distances plus courtes entre sommets. L'apprentissage le plus rapide concerne les graphes multipôles très centralisés puis les SmallWorls et graphes réguliers de faible diamètre suivis des mêmes avec un diamètre plus conséquent. Viennent ensuite les structures localement denses mais globalement dispersées et en dernier des graphes non connexes. La distance moyenne peut se comprendre comme le nombre moyen d'agents devant faire une révision dans le bon ordre pour qu'une information (exemple ou hypothèse) soit accessible par un autre agent. Le diamètre est le nombre d'agents nécessaires dans le pire des cas. Si l'un comme l'autre sont plus courts il est plus probable que les successions de révisions nécessaires à la transmission des informations soient réalisées rapidement. Il faut remarquer cependant que la roue le bipole et le tripole qui ont diamètre, densité et distance moyenne égales, diffèrent par leur

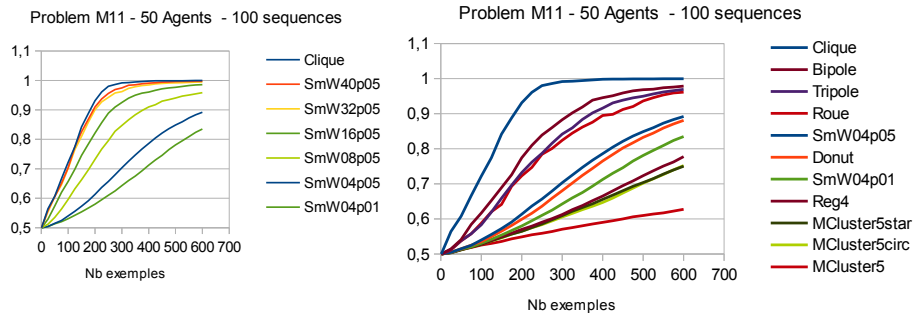


Figure 2. Évolution de la précision moyenne des hypothèses, en fonction du nombre d'exemples d'apprentissage reçus, lorsque les agents ne communiquent qu'avec leurs voisins directes. Structure similaire (gauche) : Cas des SmallWorlds de degré moyen 4 à 49. Densité fixée (droite) : Graphes de sommets de degré moyen 4 comparés à la clique

vitesse d'apprentissage. Pour une même distance entre deux sommets, l'apprentissage est favorisé en présence d'un plus grand nombre de chemins de cette longueur.

4. Importance relative de l'adoption d'hypothèse et de la mémorisation

Les graphes ne bénéficient pas tous de la même manière de l'adoption d'hypothèse et de la mémorisation d'exemples. L'analyse pour chacun des graphes de leur courbe d'apprentissage dans les deux situations d'adoption d'hypothèse (avec et sans oubli des exemples externes, Figure 3) et les deux situations sans adoption d'hypothèse (Figure 4) permet de remarquer plusieurs propriétés : (1) La propagation d'hypothèse est toujours bénéfique sur la vitesse d'apprentissage. (2) Pour des graphes de diamètre 2, il y a très peu d'influence de la mémorisation d'exemples en présence d'adoption d'hypothèses. (3) Pour des diamètres plus grand, la mémorisation d'exemples est plus déterminante à terme que l'adoption d'hypothèses. (4) Le début de la courbe d'apprentissage (100 premiers exemples) est plutôt déterminée par l'adoption d'hypothèse.

5. Conclusion

L'apprentissage est favorisé dans les réseau denses et de faible distance moyenne. Il bénéficie de la mémorisation d'exemples et de l'adoption d'hypothèses. Cette étude est préliminaire à une modélisation plus fine mettant en œuvre les stratégies de décision nécessaires aux agents dans le choix de leur hypothèse courante lorsque plusieurs hypothèses cohérentes leurs sont soumises simultanément par d'autres agents (adoption de l'une d'entre elles, construction d'une hypothèse résultante, etc ...). Ces stratégies déterminent au niveau macroscopique la formation de théories, et l'opposition entre celles-ci, dans le réseau.

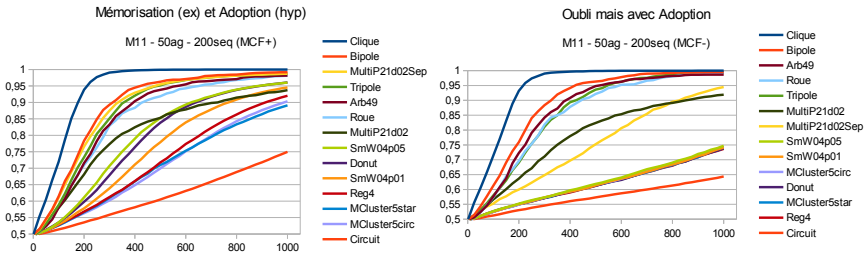


Figure 3. Évolution de la précision moyenne des hypothèses en fonction du nombre d'exemples d'apprentissage reçus ; protocoles avec adoption d'hypothèse et avec(gauche) ou sans(droite) mémorisation des exemples externes

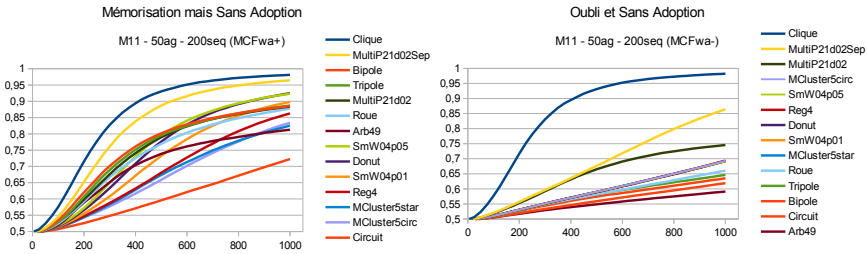


Figure 4. Évolution de la précision moyenne des hypothèses en fonction du nombre d'exemples d'apprentissage reçus ; protocoles sans adoption d'hypothèse, avec(gauche) ou sans(droite) mémorisation des exemples externes

Bibliographie

Bourgne G., Bouthinon D., El Fallah Seghrouchni A., Soldano H. (2009, novembre). Collaborative concept learning: non individualistic vs individualistic agents. In *IEEE International Conference on Tools with Artificial Intelligence (ICTAI)*, p. 653–657. Newark, USA, IEEE Computer Society.

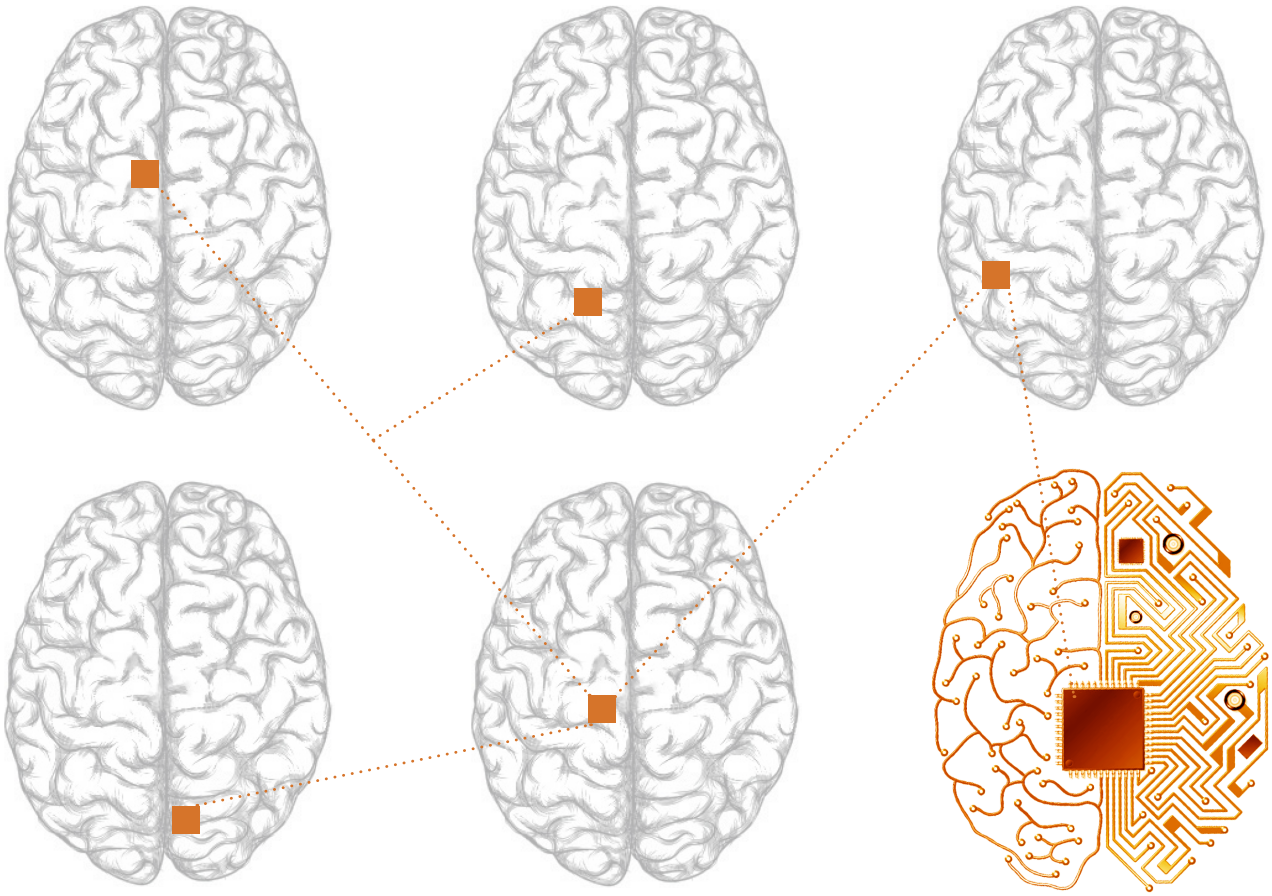
Bourgne G., El Fallah Seghrouchni A., Soldano H. (2009). Learning in a fixed or evolving network of agents. In *ACM-IAT'09*. IEEE.

Guille A., Hacid H., Favre C., Zighed D. A. (2013). Information diffusion in online social networks: a survey. *SIGMOD Record*, vol. 42, n° 2, p. 17–28.

Kevin J.S. Zollman. (2007). *Network Epistemology*. Thèse de doctorat non publiée, university of California, Irvine.

Ontañón S., Plaza E. (2010). Multiagent Inductive Learning: an Argumentation-based Approach. In J. Fürnkranz, T. Joachims (Eds.), *ICML*, p. 839–846. Omnipress.

Veillon L.-M. (2015). *Apprentissage collaboratif de proximité.* Consulté sur https://lipn.univ-paris13.fr/~veillon/



PFIA 2015

http://pfia2015.inria.fr

Plate-forme Intelligence Artificielle

Rennes du 29 juin au 3 juillet 2015