Apprentissage statistique dans les graphes et les réseaux sociaux

Patrick Gallinari

Collaboration : L. Denoyer, S. Peters

Université Pierre et Marie Curie

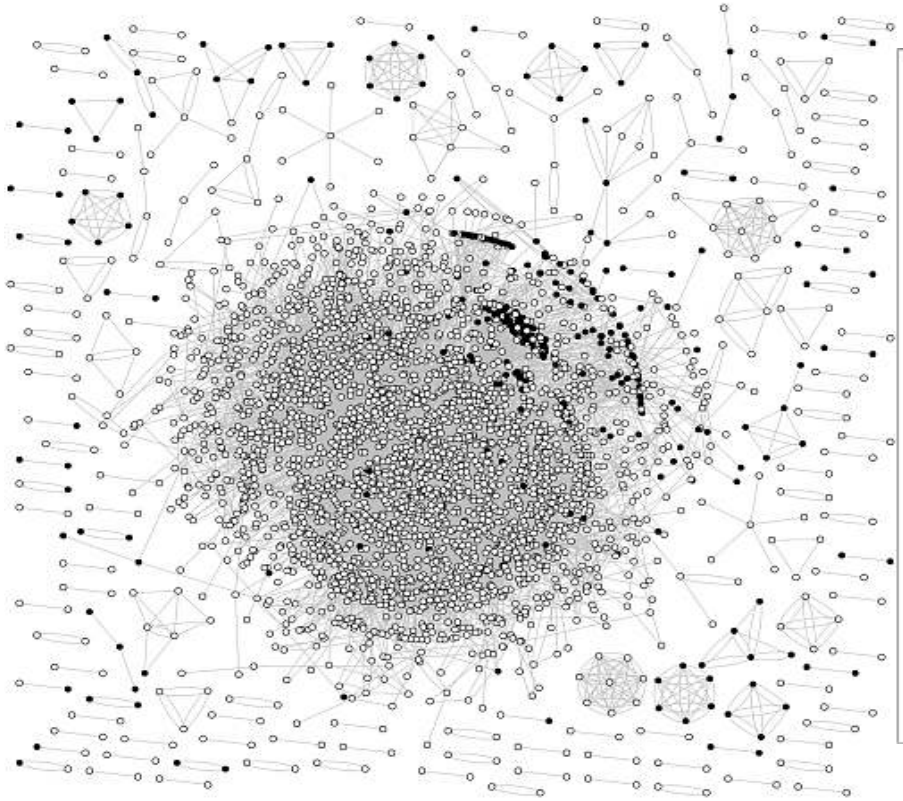
AAFD – 2010

Plan

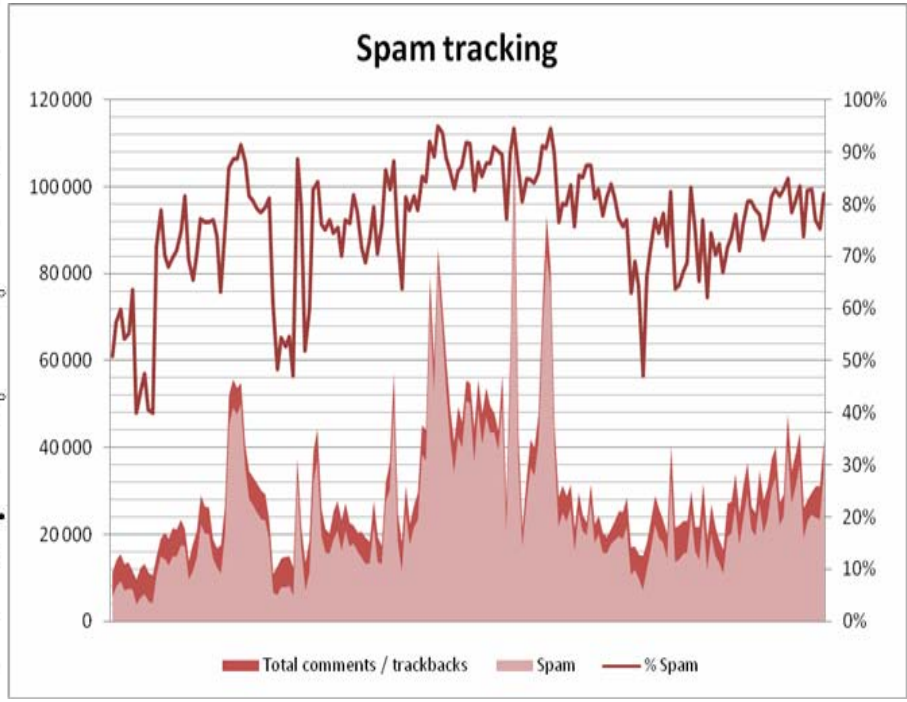
- Motivations et Problématique
- Classification dans les graphes - méthodes
  - Classification itérative
  - Apprentissage semi-supervisé
- Cas d'étude : annotation d'image
  - Classification itérative
  - Apprentissage semi-supervisé

Classification dans les graphes
Motivations – Exemple Spam

Web



Blogs

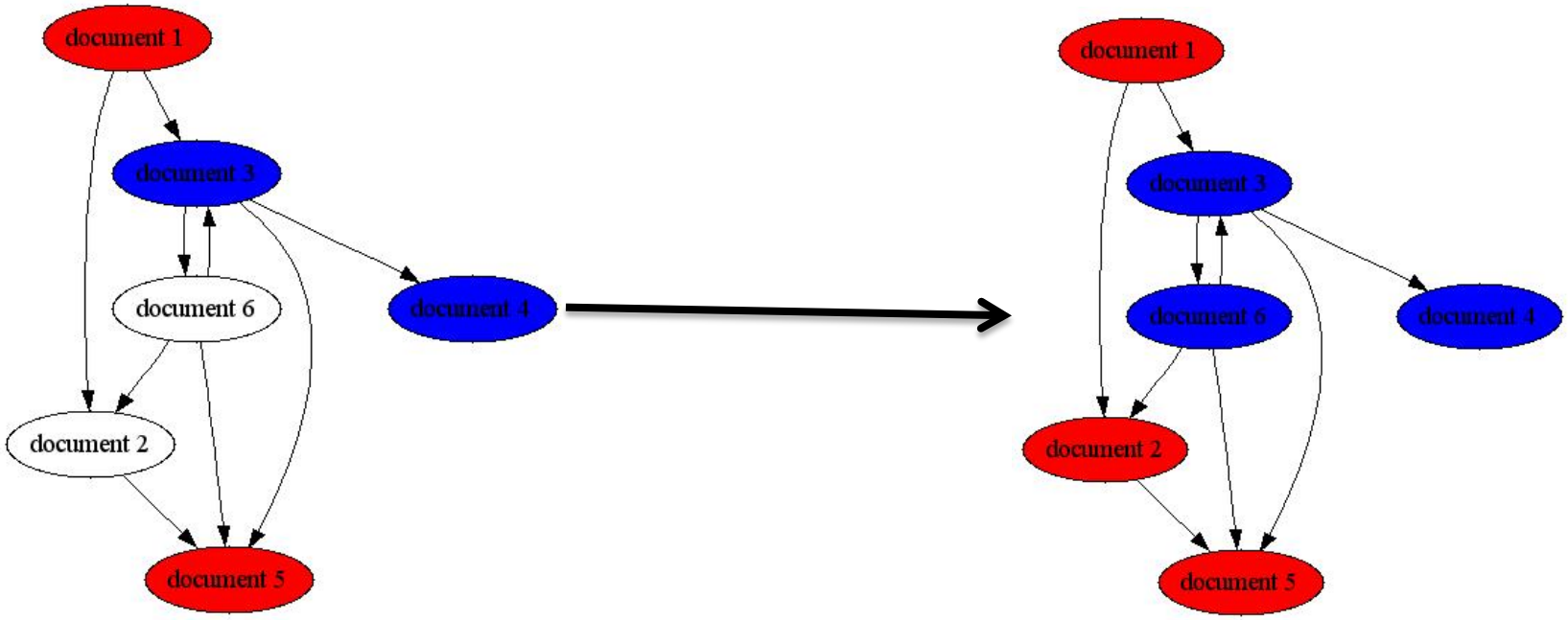


Classification dans les graphes
Motivations

- Problèmes génériques pour de nombreuses applications
  - Spam
    - Web, blogs, médias sociaux
  - Classification
    - Pages web, fichiers P2P
    - amis dans facebook
    - utilisateurs
  - Annotation
    - textes, images, vidéos,
  - Recherche d'information
    - ad hoc, distillation

Classification dans les graphes
Problématique

- Problème



Classification dans les graphes
Problématique

- Problème
  - Inférer de l'information sur des nœuds inconnus à partir de l'observation partielle des étiquettes d'un graphe
  - Utiliser les corrélations entre étiquettes
- Variété de problèmes
  - Graphes dirigés / non dirigés
  - Arcs valués / binaires
  - Relations implicite / explicite
  - Prise en compte
    - Information relationnelle
    - Information de contenu dans les nœuds
  - Graphes multi-relationnels
  - Graphes dynamiques

Classification dans les graphes
Difficultés

- Résoudre le problème d'affectation d'étiquettes est en général complexe / prohibitif
- Les algorithmes d'inférence exacts quand ils existent sont trop couteux
- Etudes
  - Variantes simplifiées du problème
  - Algorithmes d'inférence approchés

Plan

- Motivations et Problématique
- Classification dans les graphes - méthodes
  - Classification itérative
  - Apprentissage semi-supervisé
- Cas d'étude : annotation d'image
  - Classification itérative
  - Apprentissage semi-supervisé

Classification dans les graphes
Méthodes

- 2 grandes familles de méthodes
- Contexte transductif
  - Semi-supervisé
  - Méthodes de propagation d'étiquettes (Zhu et al. 2002, Belkin 2003, Zhou et al. 2003)
- Contexte inductif
  - Classification collective (Lu et al. 2003, Neville et al. 2000, Sen et al. 2009)
    - Contenu + structure
    - Méthodes itératives

Classification dans les graphes
Cadre transductif

- Propagation d'étiquettes
  - Graphe défini par la similarité des données
  - Propager les étiquettes en fonction de la similarité des données
  - Initialement uniquement de la propagation d'étiquettes sur la structure
  - Récemment prise en compte du contenu
    - E.g. Webspam – système Witch (Castillo et al. 2008)

Cadre transductif
Exploiter la consistence des données

- Utiliser simultanément l'information locale (les voisins partagent les mêmes étiquettes) et globale (structure des données)

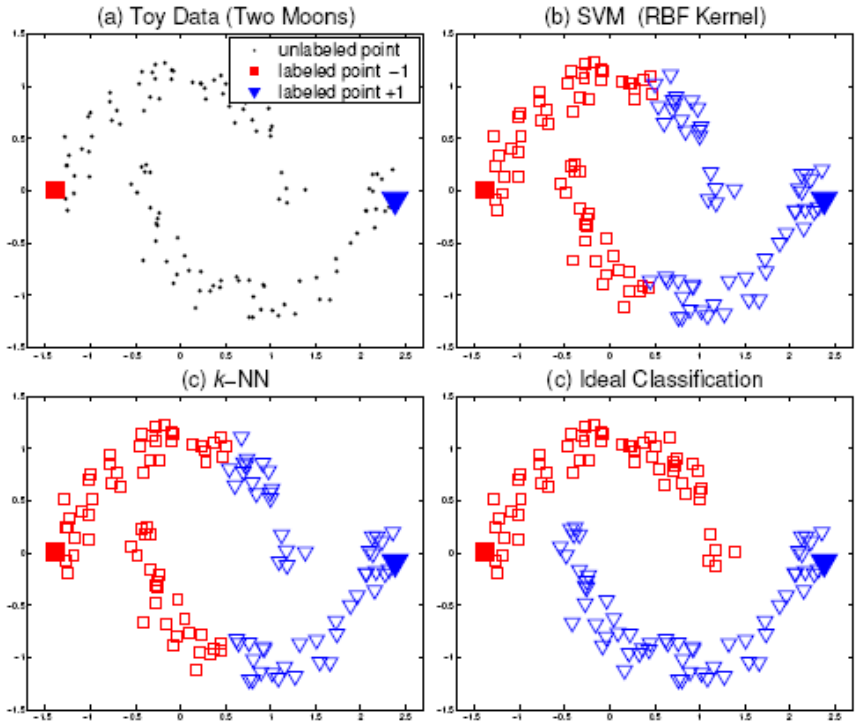


Fig. from Zhou et al. 2003

cadre transductif
Notations

- Graphe (V,E)
  - Sommets V
    - $V = V_{L(abeled} \cup V_{U(nlabeled}$
  - Arêtes E
    - E: w_{ij} poids de (i,j)
  - Etiquettes Y
    - $Y = (y_1,\ldots,y_L)$ etiquettes des sommets $(x_1,\ldots,x_L) = V_L$
- Fonction de classification
  - $F(x_i)$, $x_i \in V$

Cadre transductif
Approche générale

- Apprendre les étiquettes désirées sur V_L

$$\arg\min_F \sum_{x_i \in V_L} \Delta(F(x_i) - y_i)$$

- Tout en capturant la régularité structurelle du graphe

$$\arg\min_F \sum_{(i,j) \in E} w_{ij} (F(x_i) - F(x_j))^2$$

- Optimiser un coût **global**

$$\arg\min_F \sum_{x_i \in V_L} \Delta(F(x_i) - y_i) + \beta . \sum_{(i,j) \in E} w_{ij} (F(x_i) - F(x_j))^2 + \lambda . regularization(F)$$

Cadre transductif
Différentes instances

- Modèles de propagation purs (Zhou et al 03, 04)
  - $F(x_i) = F_i$: ne dépend pas des caractéristiques des données
    - Marches aléatoires
- Fonction de similarité générale entre les nœuds (Belkin et al. 04)

$$\arg\min_{F} \sum_{x_i \in V_L} \Delta(F(x_i) - y_i) + \beta \sum_{(i,j) \in E} K(x_i, x_j).(F(x_i) - F(x_j))^2$$

  - K function noyaux
    - En général seuillée
- Graphes dirigés, non dirigés

Classification Collective Algorithmes

- Schéma habituel
  - Bootstrap
    - Calculer une etiquette pour chaque noeud en utilisant un classifieur local
    - N'importe quel classifieur peut être utilisé
  - Itérer
    - Calculer l'etiquette du graphe en utilisant l'information contextuelle
    - Des itérations sont nécessaires, car les nouvelles étiquettes pour les noeuds dans N(i) fournissent de nouvelles informations pour y_i
    - La plupart des méthodes de classification collective nécessitent
      - Un classifieur relationnel
      - Une politique d'itérations

Classification Collective Méthodes

- Gibbs
- Iterative classification
- Relaxation labeling
- Stacked learning
- Random walks …..

Classification Collective
Vecteurs de caractéristiques

- Pour des classifieurs vectoriels, x_i doit être de taille fixe
  - Les voisinages N(i) peuvent être de taille variable pour different s noeuds i
  - Solution habituelle : utiliser des caractéristiques aggrégées pour construire des vecteurs de caractéristiques de taille fixe
    - e.g. # etiquettes pour la classe k dans N(i), fréquence relative de la classe k dans N(i), etiquettes majoritaires dans N(i), ….
    - La valeur de x_i peut changer d'une iteration à l'autre
    - Les caractéristiques x_i doit être recalculées à chaque itération

Classification Collective
ICA(Neville et al 2000, Lu et al. 2003)

- Boostrap
  - Pour chaque noeud non étiqueté i
    - Classifieur local
      - Calculer x_i
      - Calculer etiquette y_i en utilisant les noeuds connus dans N(i) : $y_i = F(x_i)$
- Iterer
  - Générer un ordre sur les noeuds non etiquetés
  - Pour chaque noeud non étiqueté i,
    - Classifieur relationnel
      - Calculer x_i
      - Calculer etiquette y_i dans le contexte N(i) : $y_i = F(x_i)$

Classification Collective
Echantillonnage de Gibbs

- Boostrap
  - Pour chaque noeud non étiqueté i
    - Classifieur local
      - Calculer x_i
      - Calculer etiquette y_i en utilisant les noeuds connus dans N(i) : $y_i = F(x_i)$
    - Pour chaque etiquette l
      - Counts [i, l] = 0
- Iterer
  - Générer un ordre sur les noueds non etiquetés
  - Pour chaque noeud non étiqueté i
    - Classifieur relationnel
      - Calculer x_i
      - Calculer etiquette y_i dans le contexte N(i) : $y_i = F(x_i)$
    - Counts [i, y_i] = Counts [i, y_i] +1
- $y_i = \text{argmax}_l \text{count}[i, l]$

Classification Collective
Stacked graphical learning

- Différence principale : phase d'apprentissage
- Idées
  - Entrainer un classifieur local y = F(x)
  - Entrainer un second classifieur en utilisant à la fois l'entrée x et les étiquettes prédites dans N(i)
  - Utiliser le stacked learning

- En général, nécessite peu (1 !) d'iterations

Classification Collective
Stacked graphical learning

- Apprentissage
  - Boostrap
    - Apprendre un classifieur local F^0 sur l'ensembtle d'apprentissage D
  - Iterer k = 1 à K
    - Construire l'ensemble d'apprentissage D^k en augmentant x_i avec $Y_{N(i)}$:
      - $x^k = (x, Y_{N(x)})$
    - Apprendre F^k sur D^k
    - Note : cette étape utilise du stacked learning
  - Modèle final : F^K
- Inférence
  - $y^0 = F^0(x)$
  - Pour k = 1 àK
    - Calculer x^k comme ci dessus
    - $y^k = F^k(x^k)$
  - $y^K = F^K(x^K)$

Classification Collective
Stacked learning

- Pour assurer la généralisation l'apprentissage est effectué par stacked learning
- Ensemble d'apprentissage D
  - Soit D_1, .., D_m une partition deD
  - F^k est. appris comme suit :
    - Entrainer m fonctions f_i
      - » f_i est appris sur $D - D_i$
      - » Soit $x \in D_i$, $y = F(x) = f_i(x)$

- Note
  - A chaque itération, une partition différente sera utilisée
  - Permet d'éviter le sur-apprentissage

Plan

- Motivations et Problématique
- Classification dans les graphes - méthodes
  - Classification itérative
  - Apprentissage semi-supervisé
- Cas d'étude : annotation d'image
  - Apprentissage semi-supervisé - transductif
  - Classification itérative - inductif

Annotation d'images

- Problématique
  - Trouver automatiquement une liste de tags pour une image
- Tâche classique
  - Très nombreuses publications
  - Différentes tâches
    - Bases professionnelles (Corel : 5 K images, 370 tags)
    - Bases annotées par contenu, concepts visuel + ontologie (Image CLEF 2009, 25 K images, 53 concepts)
    - Photos personnelles, partagées (Flickr) …
- Tâche(s) difficile(s)
  - Les performances restent faibles …
  - Cas des bases personnelles
    - les étiquettes choisies manuellement par les utilisateurs sont souvent
      - Imprécises, Ambigües, Inconsistantes, Sujettes à une grande variabilité
    - Comment étiqueter avec des étiquettes « complexes » qui ne dépendent pas directement du contenu visuel ?
    - Sites collaboratifs : très grand nombre de tags

Annotation d'images
Exemple

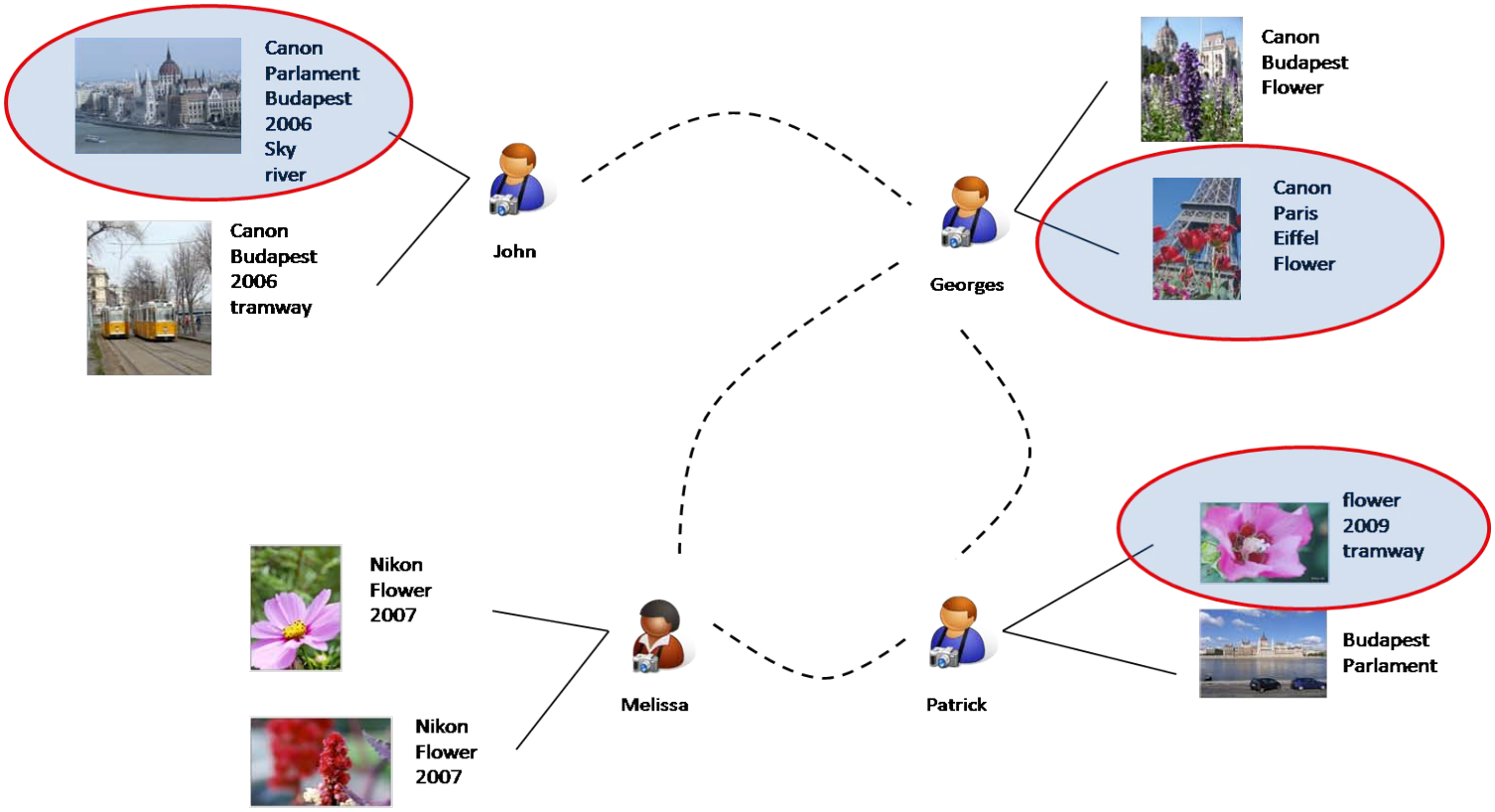
maison parlement Budapest Europe est océan rivière voiture ciel été 2006

- Etiquettes « visuelles » : Maison, Océan, Rivière, Voiture, Ciel
- Etiquettes « non visuelles » : Budapest, Europe, Été, 2006, Est
- **Idée:** Utiliser d'autres sources d'informations pour l'annotation
  - Quelques travaux en utilisant les données EXIF (GPS, Boussole, Date)
  - Utilisation de l'information relationnelle entre les images
    - Similarités (quelques travaux)
    - **Réseau social (ici)**

Annotation d'images
Méthodes

- Apprendre la correspondance directe entre contenu image et tags
  - Etiqueter des images en concepts visuels
    - SVM, perceptron, CRF, etc.
    - Modèles génératifs : HMMs et variantes, traduction, …
- Apprendre des cooccurrences entre mots visuels et textuels
  - LDA, PLSA, …
- Propagation de tags
  - Utiliser les similarités entre contenus ou meta-données pour « propager les tags » sur le graphe de similarités

Problématique



- Problème : on connait les étiquettes entourées, inférer les autres

Annotation d’images
Cadre transductif

Denoyer – Gallinari – (ICWSM 2010)

Annotation d'images
Cadre transductif

- Problème
  - généraliser un étiquetage partiel dans le réseau social
- Modèle basé sur deux idées clefs:
  - Il ordonne correctement les étiquettes des images étiquetées
  - Deux images connectées par une relation auront tendance à avoir les mêmes scores (Hypothèse de régularité)
- Exploite
  - Contenu
  - Relations

Annotation d'images
Cadre transductif

- Ensemble d'images $I = (i^1, \ldots, i^n)$
  - Ensemble étiqueté $I = (i^1, \ldots, i^l)$
  - Ensemble non-étiqueté $I = (i^{l+1}, \ldots, i^n)$
- Ensemble d'étiquettes possibles: $T = (1,\ldots,T)$
- Une description vectorielle de chaque image x^j
- y_t^j score du label t pour l'image j

- **Problématique:** retrouver les y_t^j pour les images non étiquetées

Annotation d'images - Cadre transductif
Modèle d'ordonnancement sur le contenu seul

- f_θ^{PR} fonction d'annotation paramétrée par θ
  - Coût d'ordonnancement pairwise :
    - Définit un ordre sur les étiquettes de chaque image

$$\Delta_\theta(i^k, \mathbf{y^k}) = \sum_{(t,t'):\mathbf{y}_t^k > \mathbf{y}_{t'}^k} h(f_\theta^{PR}(i,t) - f_\theta^{PR}(i,t'))$$

    - h fonction hinge
- Coût PR (Pairwise Ranking)
  - Apprendre l'espérance des ordonnancements sur les étiquettes connues

$$\mathcal{L}_{PR}(\theta) = \sum_{i^k \in \mathcal{I}_l} \Delta_\theta(i^k, \mathbf{y^k}) + \lambda_{reg}||\theta||^2$$

Annotation d'images - Cadre transductif
Prise en compte de l'information relationnelle

- On considère une information relationnelle

$$\mathcal{R} = \{w_{j,k}\}, j \times k \in [1..n]^2\}$$

  - $w_{j,k}$ est la force de la relation entre les images j et k
- Cette information relationnelle provient:
  - De relations implicites (similarités)
  - De relations explicites issues du réseau social
    - Auteurs, réseau d'amitiés,…

Annotation d'images - Cadre transductif
Modèle d'ordonnancement relationnel

- Coût pour la régularité sur information relationnelle

$$L_{REL} = \sum_{t \in T} \sum_{(j,k) \in R} w_{j,k} \left(f_\theta^{GPR}(j,t) - f_\theta^{GPR}(k,t) \right)^2$$

Fonction de scoring

- La fonction de scoring *f* dépend du contenu des image $\Phi(\mathbf{x}^k, t)$ mais aussi de la régularité induite sur le graphe:
  - Score du tag *t* pour l'image *k* :

$$f_{\theta,\xi}^{GPR}(k,t) = \langle \theta, \Phi(\mathbf{x}^k, t) \rangle + \xi_{k,t}$$
$$= f_{\theta}^{PR}(k,t) + \xi_{k,t}$$

- Les variables $\xi_{k,t}$ ont pour but de permettre au modèle de décider un score qui respecte au mieux la régularité du graphe

Annotation d'images - Cadre transductif
Modèle Graph Pairwise Ranking - GPR

$$\mathcal{L}_{GPR}^{\theta\xi} = \sum_{i^k \in \mathcal{I}_l} \sum_{(t,t'):y_t^k > y_{t'}^k} h(\langle \theta, \Phi(\mathbf{x}^k, t) \rangle + \xi_{k,t} - \langle \theta, \Phi(\mathbf{x}^k, t') \rangle - \xi_{k,t'}) \text{ (terme 1)}$$

$$+ \lambda_{REL} \sum_{t} \sum_{(j,k):w_{j,k}>0} w_{j,k}(\langle \theta, \Phi(\mathbf{x}^j, t) \rangle + \xi_{j,t} - \langle \theta, \Phi(\mathbf{x}^k, t) \rangle - \xi_{k,t})^2 \text{ (terme 2)}$$

$$+ \lambda_{reg} ||\theta||^2 \text{ (terme 3)}$$

$$+ \lambda_{slack} \sum_{k,t \in [1..\ell] \times [1..T]} \xi_{k,t}^2 \text{ (terme 4)}$$

- Terme 1 : Coût d'ordonnancement sur les images étiquetées
- Terme 2 : Régularité des scores sur la structure du graphe
- Terme 3 : Régularisation des paramètres de contenu
- Terme 4 : Régularisation des variables d'ajustement

Annotation d'images - Cadre transductif
Expériences - Flickr

Fonction caractéristique		**Description**
ψ^{image}		Histogrammes RGB normalisés avec 48 couleurs
ψ^{text}		Vecteurs TF-IDF Normalisés calculés sur les titres et les descriptions des images.
Relations		**Poids (0 si pas de relation)**
Relations Implicites	w^{image}	$w^{image}_{j,k} =< \psi^{image}(i^j); \psi^{image}(i^k) >$. La relation correspond à une similarité visuelle entre images.
	w^{text}	$w^{text}_{j,k} =< \psi^{text}(i^j); \psi^{text}(i^k) >$. La relation correspond à une similarité textuelle entre images.
Relations explicites (sociales)	w^{author}	$w^{author}_{j,k} = \begin{cases} 1 \text{ si image } j \text{ et } k \text{ ont le même auteur.} \\ 0 \text{ sinon} \end{cases}$
	$w^{friends}$	$w^{friends}_{j,k} = 1$ si les auteurs des images j et k sont amis.

Annotation d'images - Cadre transductif
Expériences Flickr

- Trois corpus issus de Flickr

Corpus	C1	C2	C3
Nombre d'images	519	801	3 183
Nombre d'auteurs	100	100	1 000
Nombre d'étiquettes	**32**	**326**	**25**
Taille des vecteurs ψ^{text}	990	990	4 460
Nombre de relations de type w^{image}	$\approx$ 120 000	$\approx$ 260 000	$\approx$ 2 millions
Nombre de relations de type w^{text}	$\approx$ 90 000	$\approx$ 140 000	$\approx$ 1.3 millions
Nombre de relations de type w^{author}	$\approx$ 9 000	$\approx$ 20 000	$\approx$ 100 000
Nombre de relations de type $w^{friends}$	$\approx$ 12 000	$\approx$ 30 000	$\approx$ 320 000

Résultats

- Relations sociales influentes
  - Auteurs, friends
- Relations de contenu peu influentes
- Classifieur par contenu seul peu performant

Résultats

Corpus :			C1			C2		
Valeur de (p) :			25%	50 %	75 %	25%	50 %	75 %
Caractéristiques	**Relations**	**Modèle**	Précision moyenne (.. %)					
ψ^{image}		PR	27	25.6	23.4	8.6	8.1	7.5
	w^{author}	GPR	**55.7**	**59.3**	**45**	**39.7**	**33.5**	**24.3**
	$w^{friends}$	GPR	51.5	49.3	42	25.6	21.3	16.6
	w^{image}	GPR	28.3	26.9	24.7	8.4	7.9	7.8
	w^{text}	GPR	29.9	26.6	24.7	8.8	8.2	8.1
ψ^{text}		PR	41.5	38.5	34.4	20.6	18.7	15.3
	w^{author}	GPR	**59.7**	56.8	51.5	41.4	**39.2**	**32**
	$w^{friends}$	GPR	59	**58.8**	**52**	**43.2**	38.9	31.5
	w^{image}	GPR	32	27.6	27.3	15.9	13.1	12.1
	w^{text}	GPR	34	35.4	34	15.6	16.8	15.4

Corpus :			C3		
Valeur de (p) :			25%	50 %	75 %
Caractéristiques	**Relations**	**Modèle**	Précision moyenne (.. %)		
ψ^{text}		PR	33.2	31.7	30.4
	w^{author}	GPR	**40.5**	36.1	33.7
	$w^{friends}$	GPR	39.1	**37.2**	**35.3**

Annotation d'images - Cadre Inductif Annotation Itérative dans des réseaux multi-relationnels

Peters – Denoyer – Gallinari – (Asonam 2010)

Annotation d'images - Cadre Inductif

- But
  - développer des modèles permettant la prise en compte de relations multiples simultanément
    - Propager les scores en utilisant l'ensemble des relations
    - Exploiter les corrélations potentielles entre les différentes relations
  - Multi-étiquettes
    - Calculer un score pour chaque étiquette possible
- Extension multi-relationnelle de l'algorithme ICA
- Originalité
  - Multi-étiquettes
  - Multi-relationnel

Annotation d'images - Cadre Inductif Algorithme

- Même forme générale que ICA
- Multirelationnel
  - Les relations entre deux nœuds i et j sont représentées par un vecteur réel avec un poids pour chaque relation

$$r_{i,j} = \begin{pmatrix} r_{i,j}^{1} \\ \cdot \\ \cdot \\ \cdot \\ r_{i,j}^{R} \end{pmatrix}$$

  - Le modèle permet plusieurs modes de propagation
    - 2 sont utilisés dans les expériences
  - Générique
    - Les poids varient selon la relation, mais sont les mêmes pour toutes les étiquettes
  - Par étiquette
    - Les poids dépendent à la fois des relations et des étiquettes

Annotation d'images - Cadre Inductif Algorithme

- Un score est calculé en chaque nœud, pour chaque étiquette

$$f_\theta(n_i, l, S^{(t)}(n_i)) = \tanh(< \theta, \Phi(n_i, l, S^{(t)}(n_i)) >)$$

- Avec
  - n_i - nœud i
  - l - étiquette
  - $S^{(t)}(n_i)$ scores des voisins
  - $\Phi(.)$ - une représentation jointe du nœud n_i , de l'étiquette et du contexte

Annotation d'images - Cadre Inductif Expériences - Flickr

- Relations
  - explicites
    - Authorship - AR
    - Friendship - FR
    - Comments - CR
    - Same month - MR
  - Implicites
    - Similarité textuelle - TR
- Small set
  - 10 K images de 1K utilisateurs, environ 1K étiquettes
- Large set
  - 47 K images, 100 étiquettes

- Expériences
  - 2 modes de propagation
  - Avec et sans contenu (texte)
  - 4 conditions expérimentales au total

Annotation d'images - Cadre Inductif
Quelques résultats - Flickr (small)

- Relationnel vs contenu seul (TC text, VC visual)

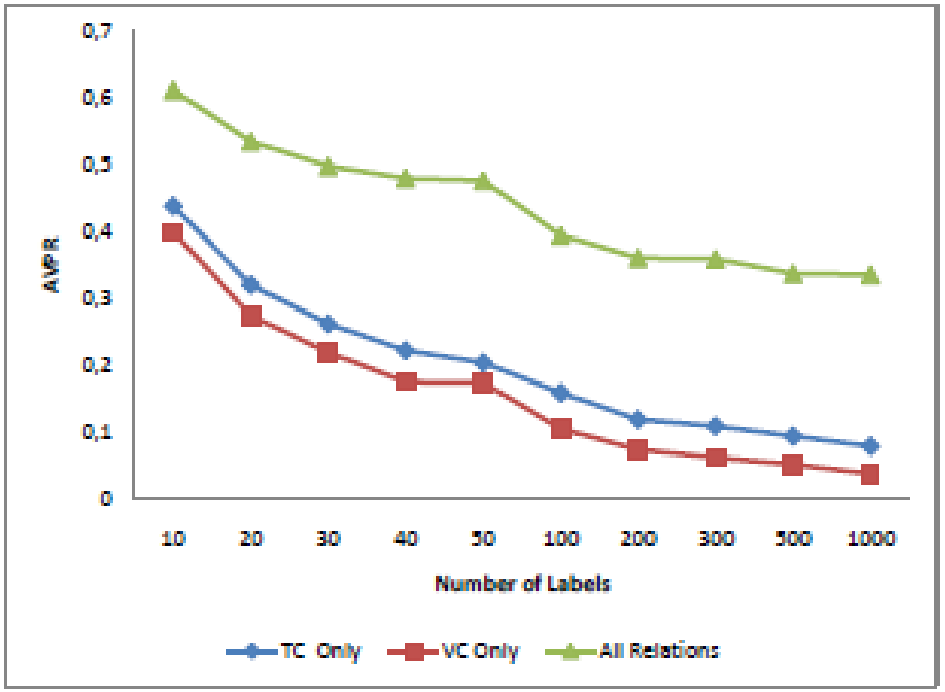


Fig. 2. AVPR for Multi-relational IMMCA with Φ_c^{GPS} propagation scheme versus Content Only models (max-margin perceptron) w.r.t the number of labels L

Annotation d'images - Cadre Inductif Quelques résultatsFlickr (small)

- Mono-Relational vs multi-relational

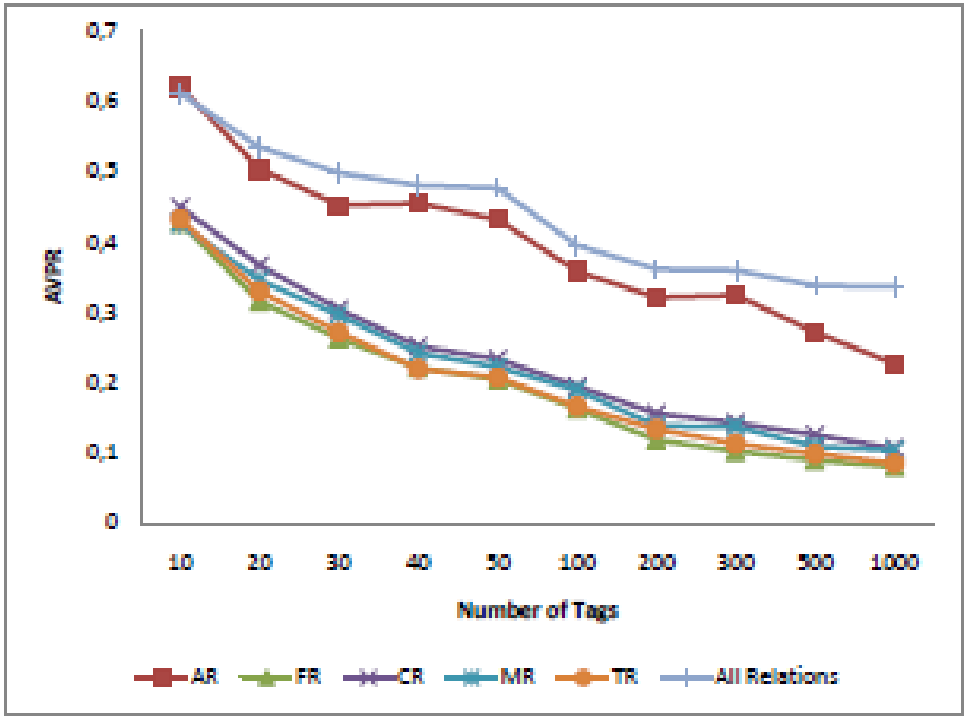


Fig. 3. AVPR for Mono-relational versus Multi-relational IMMCA with Φ_c^{GPS} propagation scheme using Textual Content for 10% in training set w.r.t the number of labels L

Annotation d’images - Cadre Inductif Quelques résultatsFlickr

- Relationnel est plus important que le contenu seul
- Multi-relationnel meilleur que mono-relationnel
- Relationnel sans contenu aussi bon que relationnel avec contenu
  - Nature très bruitée des données Flickr (contenu textuel et visuel)

Conclusion

- La majorité des études sur les données réelles issues de réseaux sociaux restent largement exploratoires
  - On décrit le phénomène observé
  - Eventuellement on le modélise
- Assez peu d'études sur les modèles prédictifs
- Beaucoup reste à faire
  - Modèles
    - E.g. prise en compte de données dynamiques
  - Compréhension des phénomènes
    - Pluri-disciplinaire

Références

- *McDowell, L. K.; Gupta, K. M.; and Aha, D. W. 2007. Cautious inference in collective classification. In Proceedings of AAAI.*
- *Neville J and Jensen D, 2000. Iterative classification in relational data. In Proceedings of the Workshop on Learning Statistical Models from Relational Data at the Seventeenth National Conference on Artificial Intelligence (AAAI), pages 13–20*
- *Neville, J., and Jensen, D. 2007. Relational dependency networks. Journal of Machine Learning Research.*
- *Zhou, D., Manavoglu, E., Li, J., Giles, C. L., and Zha, H. 2006. Probabilistic models for discovering e-communities. In Proceedings of the 15th international Conference on World Wide Web (Edinburgh, Scotland, May 23 - 26, 2006*
- *Zhou, D., O. Bousquet, T.N. Lal, J. Weston and B. Schölkopf. 2004 Learning with Local and Global Consistency. Advances in Neural Information Processing Systems 16, 321-328. (Eds.) Thrun, S., L. Saul and B. Schölkopf, MIT Press, Cambridge, MA,*
- *Denoyer L., P. Gallinari, 2010, A Ranking based Model for Automatic Aannotation in a Social Network. In ICWSM 2010*
- *Peters S., L. Denoyer, P. Gallinari, 2010, Iterative Annotation of Multi-relational Social Networks, ASONAM 2010,*