Modèles d'Ordonnancement pour l'Annotation Automatique d'Images dans les Réseaux Sociaux

Ludovic Denoyer et Patrick Gallinari

LIP6 - Université Pierre et Marie Curie
104 avenue du président Kennedy - 75016 PARIS

RÉSUMÉ. *Nous proposons un modèle d'ordonnancement de données relationnelles pour apprendre automatiquement à annoter des images dans les sites permettant le partage social d'images. Ce modèle apprend à associer une liste ordonnée d'étiquettes à une image en considérant simultanément l'information de contenu (texte/image) et les informations relationnelles entre les images. Il est capable d'utiliser aussi bien des informations relationnelles implicites comme les similarités visuelles ou les informations relationnelles explicites comme l'amitié entre deux utilisateurs, où le fait que deux images possèdent le même auteur. Il peut être utilisé aussi bien pour l'annotation d'images non-annotées ou pour aider un utilisateur à annoter ses propres images. Le modèle lui-même est basé sur un algorithme transductif qui apprend à la fois à partir des images déjà annotées et des images à annoter. Les expériences menées sur un corpus réel extrait de Flickr montrent l'efficacité de ce modèle, particulièrement quand celui-ci utilise des informations extraites du réseau social sous-jacent aux images.*

ABSTRACT. *We propose a relational ranking model for learning to tag images in social media sharing systems. This model learns to associate a ranked list of tags to unlabeled images, by considering simultaneously content information (visual or textual) and relational information among the images. It is able to handle implicit relations like content similarities, and explicit ones like friendship or authorship. It can be used either for fully automatic image labeling or for helping the user with a ranked list of candidate tags. The model itself is based on a transductive algorithm thats learns from both labeled and unlabeled data. Experiments on a real corpus extracted from Flickr show the effectiveness of this model, particularly when using authorship and friendship relations.*

MOTS-CLÉS : *Annotation d'images, Réseaux sociaux, Ordonnancement, Apprentissage automatique*

KEYWORDS: *Image Annotation, Social Networks, Ranking, Machine Learning*

1. Introduction

Nous considérons le problème de l'annotation d'images dans les grands sites de partage d'images issus du Web 2.0. Ces systèmes permettent à des utilisateurs de partager des images ou des vidéos sur le Web. Les systèmes populaires comme Flickr[1] ou Youtube[2] contiennent plusieurs milliards d'images ou de vidéos. Afin de pouvoir répondre à un besoin d'information dans de si grands sites, les données sont habituellement annotées manuellement par l'utilisation d'étiquettes (ou tags) textuelles. Ces étiquettes sont alors utilisées à travers des moteurs de recherche textuel classiques.

L'étiquetage manuel de si grandes collections est une tâche très longue et coûteuse. Qui plus est, les étiquettes choisies manuellement par les utilisateurs sont souvent imprécises, ambigues, inconsistantes et sujettes à une grande variablité ((Golder *et al.,* 2006),(Matusiak, 2006)). Les méthodes automatiques d'étiquetage dont l'objectif est d'améliorer la qualité des étiquettes ont fait l'objet récemment d'une recherche intense. Cependant, à l'heure actuelle, les résultats obtenus par ces méthodes sont souvent décevants.

Les approches les plus communes considèrent le problème de l'étiquetage automatique comme un problème semi-supervisé dans lequel des informations visuelles (les images) doivent être associées à un ensemble d'étiquettes connues a priori ((Hironobu *et al.,* 1999), (Chang *et al.,* 2003), (Li *et al.,* 2008),...). Les caractéristiques utilisées par ces modèles sont le plus souvent des caractéristiques visuelles comme des histogrammes pour la recherche sur le Web ou bien alors issues de traitements d'images plus complexes pour des applications spécifiques. Quelques méthodes non-supervisées exploitant principalement des co-occurences entre caractéristiques visuelles et étiquettes ont aussi été proposées ((Barnard *et al.,* 2003)). D'autres méthodes sont basées sur la modélisation de propagation d'étiquettes : à partir d'un étiquetage initial, les étiquettes sont propagées vers les images non-annotées. Cependant, ces dernières méthodes n'utilisent souvent aucune information visuelle et juste une structure relationnelle entre les images.

Toutes ces méthodes sont assez limitées. Les méthodes de classification ou bien les méthodes à base de variables latentes par exemple ne peuvent apprendre qu'une correspondance entre étiquettes et information visuelle et ne peuvent donc pas être utilisées avec des étiquettes plus abstraites. Par exemple, en considérant l'image de la figure 1, les étiquettes *ciel*, *océan* and *rivière* peuvent certainement être inférées à travers l'utilisation d'informations visuelles tandis que les étiquettes *Budapest*, *Europe*, *est* et *2006* ne le peuvent certainement pas. De plus, ces méthodes ne peuvent pas utiliser différentes sources d'informations comme les informations de géo-localisation, la date, l'auteur ou d'autres méta-données que les systèmes existants stockent conjointement aux images. Enfin, elles n'ont pas été conçues pour exploiter les informations riches contenues par les sites "2.0" comme par exemple l'information de réseau d'amitié ou d'intérets d'utilisateurs.

1. http ://www.flickr.com
2. http ://www.youtube.com

Très récemment, quelques méthodes ont été développées pour l'exploitation d'informations relationnelles dans les collections d'images. Par exemple, Tong et al. dans (Tong *et al.,* 2006) présentent une méthode semi-supervisée permettant la propagation d'étiquettes le long de relations. De même, Cao et al. dans (Cao *et al.,* 2008) considèrent les meta-données ainsi que des modèles de propagation. Les deux systèmes cependant n'exploitent que des informations relationnelles implicites calculées à travers des similarités visuelles entre les images ou des similarités entre les étiquettes.

Nous proposons ici un nouveau modèle d'apprentissage relationnel spécialement conçu pour l'étiquetage automatique dans les sites "sociaux". Ce modèle est capable d'exploiter à la fois l'information visuelle ou textuelle contenue dans les données à annoter, ainsi que des informations relationnelles entre les données. Ces informations relationnelles peuvent être aussi bien implicites et correspondre à des similarités visuelles/textuelles que explicites à travers des relations extraites du réseau social sous-jacent comme les réseaux d'amitié par exemple.

Ce modèle est utilisé ici pour l'ordonnancement d'étiquettes pour l'annotation d'images dans les grandes collections d'images du Web. Plus précisément, il considère le problème transductif où une collection contiennent à la fois des images annotées manuellement et des images non annotées, ainsi que des relations issues du réseau social d'utilisateurs et des métadonnées. Le but est alors de fournir, pour les images non annotées, une liste **ordonnée** d'étiquettes et éventuellement de permettre aussi l'enrichissement des annotations sur les images déjà annotées.

Dans les expériences présentées ici, nous utilisons l'information issue d'auteurs ainsi que l'information d'amitié entre utilisateurs qui sont accessibles à travers les données issues de Flickr. Cependant le modèle n'est pas spécifique à ces informations et peut être utilisé avec informations ou métadonnées éventuellement disponibles. A notre connaissance, ce modèle est le premier à utiliser simultanément l'information de contenu, les métadonnées et l'information relationnelle pour l'apprentissage d'annotations.

Les contributions de l'article sont les suivantes :

– Nous proposons un nouveau modèle d'apprentissage dans les graphes permettant d'apprendre à ordonner des étiquettes en considérant des informations relationnelles entre les données.

– Le modèle est capable d'exploiter les différentes variétés d'informations présentes dans les sites sociaux

– Toutes les sources d'informations peuvent être intégrées pendant l'apprentissage à travers l'optimisation d'une fonction objective globale.

– La méthode proposée est validée sur un corpus issu du site Flickr.

De plus, le sytème est basé sur un algorithme d'ordonnacement qui peut être utilisé indifféremment pour l'annotation de nouvelles images, mais aussi pour la suggestion d'étiquettes et de mots-clefs à un utilisateur souhaitant enrichir ses propres annotations.

Tags : maison parlement Budapest Europe est océan rivière voiture ciel été 2006

Figure 1. *Une image et les étiquettes associées*

Le papier est organisé comme suit : dans la section 2 nous introduison un algorithme pour l'apprentissage d'ordonnancement d'étiquettes sur des images en ne considérant que l'information de contenu. Cette algorithme sert ensuite de "brique de base" pour le modèle présenté en section 3. Ce modèle permet l'apprentissage d'ordonnancement pour des données organisées sous forme de graphe, où les noeuds correspondent ici à des images, et les liens à des relations entre images. Dans la partie expérimentale 4, nous présentons les résultats obtenus sur différentes collections et comparons notre modèle au modèle de référence "non-relationnel" présenté ici. Ces expériences permettent de comparer les performances obtenues en considérants différentes informations de contenu ou relationnelles. Elles montrent particulièrement que, pour les collections utilisées ici, le modèle obtient les meilleures performances quand il utilise des informations extraites directement du réseau social sous-jacent aux images à annoter.

2. Modèle d'ordonnancement d'étiquettes

2.1. *Notations et Définitions*

Dans la suite, les vecteurs sont notés en gras ; les indices correspondent aux composantes d'un vecteur tandis que les exposants correspondent à des index. Par exemple, $\mathbf{x}_k^j$ correspond à la k-ème composante du j-ème vecteur $\mathbf{x}$. Nous considérons :

– Un ensemble d'images présent dans le système : $\mathcal{I} = (i^1, ..., i^n)$ où n correspond au nombre d'images.

– Une fonction caractéristique $\psi : \mathcal{I} \rightarrow \mathcal{R}^N$. Cette fonction transforme une image en un vecteur de caractéristiques de dimension N. Différentes fonctions ψ seront définies dans la partie 4. Par souci de simplicité, nous écrirons $\psi(i^j)$ sous la notation $\mathbf{x}^j$ dans la suite de l'article. $\mathbf{x}_k^j$ correspond donc à la k-ème composante du vecteur de

caractéristiques issus de l'image i^j.

– L'ensemble des étiquettes possibles est noté $\mathcal{T} = (1, .., T)$ où T correspond au nombre des étiquettes.

Nous définissons une fonction f d'ordonnancement comme :

$$\begin{aligned} f : [1..n] \times \mathcal{T} &\rightarrow \mathcal{R} \\ (j, t) &\mapsto f(j, t) = \mathbf{y}_t^j \end{aligned} \quad [1]$$

où $\mathbf{y}_t^j$ correspond au score de l'étiquette t pour l'image i^j. Plus le score est élevé, plus l'étiquette est pertinente.

L'ensemble des images $\mathcal{I}$ est composé de deux sous-ensembles :

– Un ensemble étiqueté $\mathcal{I}_l = (i^1, ..., i^\ell)$ composé de ℓ images et de leurs annotations $(\mathbf{y}^1, ..., \mathbf{y}^l)$ fournies par les utilisateurs telles que $\mathbf{y}_k^j = 1$ si k est une étiquette de l'image i^j et 0 sinon

– Un ensemble non-étiqueté $\mathcal{I}_u = (i^{\ell+1}, ..., i^{\ell+u})$ de taille u qui correspond aux images que les utilisateurs n'ont pas annotées[3].

L'annotation automatique consiste à calculer, à partir de l'ensemble d'images $\mathcal{I}$ et des annotations $(\mathbf{y}^1, ..., \mathbf{y}^l)$, un score d'ordonnancement $f(i^j, t)$ pour toutes les images de $\mathcal{I}_u$[4] et toutes les étiquettes possibles.

Afin de décrire notre modèle d'odonnancement relationnel, nous procédons en deux étapes. Dans la section suivante, nous présentons un modèle d'ordonnancement simple qui ne considère que le contenu des images et pas les relations. Ce modèle sera utilisé dans la partie expérimentale comme un modèle de référence. Il constitue une "brique" du modèle relationnel présenté en partie 3 et permet ainsi de mieux comprendre l'originalité de notre approche.

2.2. *Modèle d'ordonnancement de paires pour l'annotation basée sur le contenu*

2.2.1. *Problème d'apprentissage*

Plusieurs algorithmes d'ordonnancement ont été proposés dans la communauté d'apprentissage automatique, et, récemment, leur utilisation est devenue populaire pour la Recherche d'Information. L'algorithme décrit ici est formellement similaire à l'agorithme RankSVM présenté dans (Joachims, 2002). Nou le présentons, car il servira d'une part de modèle de comparaison, et parce qu'il constitue une brique élémentaire de notre approche. Notez que, dans notre cas, contrairement aux travaux sur

3. Notez que $\ell + u = n$

4. Cette configuration où les éléments annotés et les éléments non-annotés sont connus au moment de l'entrainement est connue sous le nom d'apprentissage transductif. L'application considérée ici justifie cette configuration. Cependant, des variantes du modèle présenté ici peuvent tout à fait être développées pour des configurations inductives plus classiques.

ce modèle, nous entrainerons ici l'algorithme à l'aide d'une procédure de descente de gradient là où les auteurs de cette méthode utilisent plutôt des méthodes d'optimisation convexe.

Nous définissons le coût ou risque d'ordonnancement de paires pour un modèle de paramètres θ pour une image étiquetée $i^k \in \mathcal{I}_l$ dont les annotations sont $\mathbf{y}^k$ ainsi :

$$\Delta_\theta(i^k, \mathbf{y^k}) = \sum_{(t,t'):\mathbf{y}^k_t > \mathbf{y}^k_{t'}} h(f^{PR}_\theta(i,t) - f^{PR}_\theta(i,t')) \quad [2]$$

où $h(.)$ correspond à la fonction *hinge loss* classique défine par :

$$h(x) = \max(0, 1-x) \quad [3]$$

$f^{PR}_\theta(i,t)$ est la fonction d'annotation définie dans l'équation 1. Ce risque correspond à une borne supérieure de l'erreur sur les paires d'exemples à ordonner et sa minimisation permet de minimiser efficacement l'erreur d'ordonnancement et particulièrement la précision moyenne (voir section 4).

Nous introduisons une régularisation en norme L2 avec un hyper-paramètre λ_{reg} [5] afin d'éviter le sur-apprentissage. La fonction objective d'ordonnancement sur l'ensmeble des images annotées s'écrit alors ainsi :

$$\mathcal{L}_{PR}(\theta) = \sum_{i^k \in \mathcal{I}_l} \Delta_\theta(i^k, \mathbf{y^k}) + \lambda_{reg}||\theta||^2 \quad [4]$$

Notez que seules les images annotées sont considérées à cette étape. Le problème d'apprentissage ainsi défini correspond à trouver les paramètres θ^* qui minimisent le coût empirique :

$$\theta^* = \text{argmin}_\theta \, \mathcal{L}_{PR}(\theta) \quad [5]$$

Ce modèle est noté **PR** par la suite (PR = Pairwise Ranking).

2.2.2. *Fonction d'ordonnancement*

Nous utilisons ici une fonction d'ordonnancement linéraire $f_\theta(k,t)$ afin de calculer le score de l'image i^k pour l'étiquette t :

$$f^{PR}_\theta(k,t) = \langle \theta, \Phi(\mathbf{x^k}, t) \rangle \quad [6]$$

où $\Phi(\mathbf{x^k}, t)$ est un vecteur de caractéristique décrivant conjointement une image i^k et une étiquette t. $\langle .,. \rangle$ correspond au produit scalaire classique. La fonction Φ utilisée ici est la fonction caractéristique multiclasse définie dans (Har-Peled *et al.*, 2002) par :

$$\Phi(\mathbf{x^k}, t) = \begin{pmatrix} (0 \cdots 0) \cdots \mathbf{x^k} \cdots (0 \cdots 0) \end{pmatrix} \quad [7]$$

5. λ_{reg} est habituellement fixé par validation croisée

$\Phi(\mathbf{x^k}, t)$ est un vecteur creux composé du sous-vecteur $\mathbf{x^k}$ à la position $N * t$ dans $\mathcal{R}^{N*T}$ et de 0 partout ailleurs. Cette notation permet d'exprimer la fonction d'ordonnamcenemt comme une simple fonction linéaire définie sur l'espace conjoint des images et des étiquettes. Elle est équivalente à l'apprentissage d'un modèle linéaire par étiquette possible, ce qui revient à T modèles linéaires indépendants en un.

3. Annotation dans un Réseau Social

Dans cette partie, nous présentons un nouveau modèle d'ordonnacement appelé GPR - pour *Graph Pairwise Ranking* - conçu pour étiqueter des données (des images ici) à l'aide de leur information de contenu, mais aussi à l'aide d'information externes comme des méta-données, ainsi que des informations relationnelles, notament extraites d'un réseau social.

Le modèle GPR est donc un modèle d'ordonnancement qui considère à la fois le contenu des images et une structure de graphe définie sur ces images . Ce graphe peut être extrait directement des images ou du texte associé à celles-ci comme dans (Tong *et al.*, 2006) (relations implicites) ou bien construit à l'aide d'un réseau social sous-jacent aux images (relations explicites). Dans la partie 4, nous explorons l'influence des différents types de relations. Dans la suite de cette partie, nous ne considérons qu'une unique relation générale.

Soit $\mathcal{R}$ un ensemble de relations entre images. Les éléments de $\mathcal{R}$ sont des scalaires : $\mathcal{R} = \{w_{j,k}\}, j \times k \in [1..n]^2\}$ où $w_{j,k} > 0$ est le poids de la relation entre l'image j et l'image k. Le couple $(\mathcal{I}, \mathcal{R})$ correspond donc à un graphe pondéré d'images.

3.1. *Modèle proposé*

Le modèle GPR est basé sur deux hypothèses. Il considère qu'un bon modèle d'ordonnancement est à la fois :

– Un modèle qui ordonne correctement les images de $\mathcal{I}_l$. (Hypothèse classique d'ordonnancenemt)

– Un modèle qui exploite la structure du graphe afin d'extraire des régularités sur la collection d'images. En particulier, nous allons exploiter les relations qui renforcent les similarités entre listes ordonnées d'étiquettes. (Hypothèse de régularité)

Si l'on considère que la structure de graphe est extraite d'un réseau d'intéret ou d'amitié par exemple, savoir que deux utilisateurs sont amis va probablement nous fournir une indication de regularité entre leurs images et les listes d'étiquettes associées - typiquement, deux utilisateurs partageant les mêmes centres d'intéret vont avoir tendance à utiliser les mêmes étiquettes. Il en sera de même si les relations considérées sont des relations de similarité visuelle entre images. L'hypothèse 1 est capturée par

la fonction objective du modèle PR présentée précédemment tandis que l'hypothèse 2 de régularité est introduite ici à travers l'ajout d'un terme relationnel :

$$\begin{aligned}\mathcal{L}_{GPR}(\theta) = &\sum_{i^k \in \mathcal{I}_l} \Delta_\theta(i^k, \mathbf{y^k}) + \lambda_{reg}||\theta||^2 \\ &+ \lambda_{REL} \sum_{t\in[1..T]} \sum_{(j,k)\in[1..n]^2} w_{j,k}\big(f_\theta^{GPR}(j,t) - f_\theta^{GPR}(k,t)\big)^2 \\ =&\mathcal{L}_{PR}(\theta) + \lambda_{REL}\mathcal{L}_{REL}(\theta)\end{aligned} \quad [8]$$

où $f_\theta^{GPR}(k,t)$ est la fonction score du modèle GPR de paramètres θ.

Le terme $\mathcal{L}_{REL}(\theta)$ vise à forcer une similarité entre les scores des étiquettes de deux images connectées, propotionnellement au poids de la connection. Une telle hypothèse de régularité est une chose classique pour les modèles de graphe en apprentissage ((Abernethy *et al.*, 2008) par exemple). Cependant, cette hypothèse n'a jamais été faite précédemment dans le cadre de l'apprentissage de listes ordonnées d'étiquettes.

3.1.1. *Fonction d'ordonnacement du modèle GPR*

Dans le modèle GPR, nous souhaitons que le score d'une image dépende de différentes sources d'informations disponibles. Principalement, le score d'une étiquette pour une image dépend de :

– l'information de contenu de l'image

– les informations relationnelles entre les images

En considérant cela, nous définissons la fonction de score comme :

$$\begin{aligned}f_{\theta,\xi}^{GPR}(k,t) &= \langle\theta, \Phi(\mathbf{x}^k, t)\rangle + \xi_{k,t} \\ &= f_\theta^{PR}(k,t) + \xi_{k,t}\end{aligned} \quad [9]$$

où $\theta \in \mathcal{R}^N$ corresponds aux paramètres d'une fonction de score basée sur le contenu uniquement et, $\xi \in \mathcal{R}^{[1..n]\times[1..T]}$ sont des variables d'ajustement additionnelles. Il y a une variable d'ajustement par image et par étiquette. En considérant n images et T étiquettes, le modèle GPR possède donc $n * T$ degrés de libertés additionnels par rapport au model PR. Ils vont permettre au modèle d'ajuster les scores des étiquettes d'une image en fonction des scores des étiquettes des images voisines dans le graphe. Tous les paramètres θ et ξ de la fonction score vont être appris simultanément comme décrit dans la partie suivante.

3.1.2. *Fonction Objective du modèle GPR*

En insérant la fonction de score décrite précédemment dans la fonction objective 8, nous obtenons la fonction objective suivante :

$$\begin{aligned}\mathcal{L}_{GPR}^{\theta\xi} = & \sum_{i^k \in \mathcal{I}_l} \sum_{(t,t'):y_t^k > y_{t'}^k} h(\langle\theta, \Phi(\mathbf{x}^k, t)\rangle + \xi_{k,t} - \langle\theta, \Phi(\mathbf{x}^k, t')\rangle - \xi_{k,t'}) \text{ (terme 1)} \\ & + \lambda_{REL} \sum_t \sum_{(j,k):w_{j,k}>0} w_{j,k}(\langle\theta, \Phi(\mathbf{x}^j, t)\rangle + \xi_{j,t} - \langle\theta, \Phi(\mathbf{x}^k, t)\rangle - \xi_{k,t})^2 \text{ (terme 2)} \\ & + \lambda_{reg}||\theta||^2 \text{ (terme 3)} \\ & + \lambda_{slack} \sum_{k,t \in [1..\ell]\times[1..T]} \xi_{k,t}^2 \text{ (terme 4)}\end{aligned} \quad [10]$$

où :

– le terme 1 correspond au coût d'ordonnancement comme défini en partie 2 où $f_{\theta,\xi}^{GPR}$ a été substitué à f_θ^{PR},

– le terme 2 correspond au terme de régularité de graphe introduit dans la section 3.1,

– le terme 3 est le terme de regularisation sur le paramètres de contenu θ,

– le terme 4 est le terme de regularisation sur les variables d'ajustement. λ_{slack} est un hyper paramètre fixé par validation croisé ou mannuellement qui penalise les valeurs élevée des variables d'ajustement et encourage les scores finaux à être proche des score basés sur le contenu seul $f_\theta^{PR}(k,t)$. Typiquement, si $\lambda_{slack} = +\infty$, $\forall k, t, \xi_{k,t} = 0$ et le modèle GPR va alors ordonner les étiquettes en ne considérant que le contenu des images. Dans ce cas, ce modèle sera utilisable sur de nouvelles images non connues au moment de l'apprentissage.

La fonction objective est minimisée selon une méthode de descente de gradient classique simultanément sur les paramètres θ et ξ. La complexité du modèle est $O(R.T.N)$ où R est le nombre d'arcs non nuls dans le graphe des images. Le modèle sera donc plus rapide pour les relations creuses (comme les relations d'amitiés) et plus long à apprendre pour les relations denses comme les similarités entre images. Cependant, les expériences montrent que les relations creuses sont souvent plus compétitives que les relations trop denses.

4. Expériences

Les expériences ont été faites sur différents corpus extraits du site Flickr[6]. Chaque corpus est composé d'un ensemble d'images et d'utilisateurs. Chaque image est dé-

6. http ://www.flickr.com

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.

Tableau 1. *Fonctions caractéristiques et relations*

crite par son contenu visuel (.jpg files) et par des meta-informations (titre, description, author, date, commentaires,....). Chaque utilisateur dans le réseau social est déccrit par son identifiant, son ensemble d'images et son ensemble d'amis. Afin d'obtenir V images et U utilisateurs, les collections ont été collectés à travers l'API Flickr [7] de la manière suivante :

– Nous avons collecté U utilisateurs dans un réseau d'amitiés : les utilisateurs sont tous amis, amis d'amis, amis d'amis d'amis,...

– Pour chaque utilisateur, nous avons collecté $\frac{V}{U}$ de ses images au hasard.

Toutes les images sont associées à un ensemble d'étiquettes définies manuellement par les utilisateurs. Pour les besoin expérimentaux, les collections ont été divisées en deux sous-ensembles : un ensemble étiqueté et un ensemble non-étiqueté (les étiquettes de ces images sont connues mais ne sont pas utilisées par l'algorithme). L'entrainement est effectué à l'aide de ces deux sous-ensembles et les étiquettes sont prédites sur les images non étiquetées. Enfin, les performances sont évaluées en fonction des étiquettes prédites et des étiquettes réelles des images non étiquetées. Nous avons utilisé 3 jeux de données. Pour chaque collection, nous avons extrait deux types de vecteur de caractéristiques décrivant les images :

– ψ^{image} correspond au contenu visuel des images.

– ψ^{text} correspond au contenu "textuel" des images extrait à partir des méta-donnée disponible sur Flickr

7. http ://www.flickr.com/services/api/

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 typew^{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

Tableau 2. *Statistiques concernant les jeux de données.*

Les fonction caractéristiques sont décrites en table 1. Nous avons délibérement utilisé de simples histogrammes pour décrire le contenu des images : en effet, nous nous intéressons ici à l'importance de considérer des informations relationnelles entre images et notamment des informations sociales et non pas au performances dans l'absolu d'un modèle de classification de contenu visuel.

Nous avons extrait 4 types de relations différentes entre images décrites dans la table 1. Deux de ces relations (w^{text} and w^{image}) sont des relations implicites qui correspondent à des similarités de contenu entre image comme fait dans (Tong *et al.*, 2006). Les deux autres (w^{author} and $w^{friends}$) correspondent à des informations issues du réseau social de Flickr (table 1). Les relations implicites qui sont très denses ont été seuillées afin de ne garder que les relations les plus fortes et ainsi réduire le nombre d'arcs dans le graphe des images.

Enfin, nous avons extrait les étiquettes les plus fréquentes et gardé les images étiquetés avec au moins l'une de ces étiquettes. Nous obtenons finalement 3 jeux de données différents décrits dans le tableau 2. Le corpus C1 possède un petit nombre d'images et d'étiquettes. Le corpus C2 possède un grand nombre d'étiquettes possibles et est utilisé pour mesurer la capacité du modèle à gérer des tâches complexes d'étiquetage. Le corpus C3 possède un grand nombre d'images et permet de mesurer la capacité du modèle à passer à l'échelle.

Ces jeux de données contiennent des étiquettes hétérogènes et l'annotation automatique est une tâche difficile. Par exemple, parmi les étiquettes conservées, nous pouvons voir les étiquettes suivantes : *beach, california, water, 2008, canon, nyc, flower, 2007, ... 2007,2008,nyc* et *canon*[8] sont quasiment impossibles à trouver en utilisant uniquement l'information visuelle des images.

8. *canon* correspond à la marque de l'appareil qui a pris la photo.

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**

Tableau 3. *Performances des modèles. La valeur de p (e.g. 25%) indique la propotion d'images non étiquetées.*

4.1. *Méthodologie*

Nous avons effectué les expériences avec différentes valeurs des hyper-paramètres à la fois pour les modèles PR et GPR. Ces hyper-paramètres sont :

– le nombre d'itérations de la descente de gradient,

– le pas de gradient,

– la proportion d'images non étiquetées notée p avec $p = \frac{u}{l+u}$,

– et les hyper-paramètres de régularisation λ_{reg}, λ_{REL} et λ_{slack}.

Nous avons lancé 3 runs pour chaque configuration de paramètres. Nous avons calculé, pour chaque image, la précision moyenne (Average Precision) de la liste ordonnée d'étiquettes renvoyée par notre algorithme. Nous reportons ici les résultats moyennés sur toutes les images et tous les runs. Nous avons enfin comparé les résultats obtenus pour différents combinaisons de caractéristiques de contenu et de relations en utilisant, comme modèle de référence, le modèle PR. Cela correspond à plusieurs milliers d'expériences, et seuls les résultats les plus significatifs sont reportés dans cet article.

4.2. *Résultats*

Le tableau 3 présente la précision moyenne obtenu sur les trois jeux de données C1, C2 et C3. Notez que, pour le corpus C3 qui est le plus grand, nous présentons uniquement les résultats issus des meilleures combinaisons de caracteristiques et de relations. Pour les corpus C1 et C2, nous avons effectué 500 itérations d'apprentissage et 3000 pour le corpus C3. Les meilleurs résultats ont été obtenus avec les va-

leurs $\lambda_{reg} = 0.1, \lambda_{slack} = 1$ et $\lambda_{REL} = 0.1$ pour les relations explicites (sociales) et 0.01 pour les relations implicites basés sur le contenu (image et texte).

Tout d'abord, nous pouvons constater que les caractéristiques textuelles (ψ^{text}) sont plus pertinentes pour annoter des images et permettent de meilleurs résultats que les caractéristiques visuelles (ψ^{image}). Cela est du au fait que les mots fournissent plus de "sémantique" que la couleur des pixels et sont plus pertinents pour retrouver des étiquettes plus abstraites comme une date ou un lieu par exemple qui ne peuvent être inférés en utilisant une description visuelle de l'image uniquement. Ensuite, l'utilisation de relations implicites ne semble pas clairement améliorer les performance du modèle, voire même peut les dégrader. Par exemple, pour le corpus C1 avec 25% d'images non étiquetées, l'utilisation de ces relations augmente la précision moyenne de $2,9\%$ si l'on utilise des caractéristiques visuelles et des relations de similarité textuelles tandis que cela dégrade les performances de $7,5\%$ si l'on utilise des caractéristiques textuelles et des relations textuelles. Enfin, l'utilisation de relations explicites issues du réseau social permet clairement d'améliorer la qualité des résultats retournés par le modèle. Pour le corpus C1 par exemple, l'utilisation de caractéristiques visuelles et de la relation d'auteur permet de multiplier la performance du modèle de référence par deux. Cet effet est particulièrement marqué pour le corpus C2 qui contient beaucoup d'étiquettes possibles et pour lequel l'annotation est plus complexe. Dans ce cas, la performance atteind 40% en utilisant les relations sociales quand le modèle de référence n'obtient que 8%. Quand on utilise des caractéristiques textuelles, cet effet est moindre mais significatif. L'augmentation de performances obtenue sur le plus gros des corpus (C3) est aussi relativement significative (de 5% à 7%).

5. Etat de l'art

Une grande variété d'algorithmes d'apprentissage on été développés pour l'annotation (semi-)automatique d'images. Les méthodes supervisées (noyaux, plus proches voisins, modèles graphiques, etc) ont été utilisées dans plusieurs expérimentations ((Frome *et al.,* 2007)). Du point de vue non-supervisé (découverte automatique d'étiquettes), les modèles à variables latentes semblent les plus intéressants avec des références populaires et anciennes comme (Barnard *et al.,* 2003) par exemple. Plus récemment, plusieurs auteurs ont tenté d'exploiter les corrélations entres caractéristiques visuelles, ou textuelles, afin d'améliorer les performances des sytèmes plus anciens, principalement à travers des méthodes de propagation d'étiquettes entre les images similaires (Cao *et al.,* 2008), (Tong *et al.,* 2006). Un algorithme d'ordonnancement dans les graphes bi-partite pour l'étiquetage de textes a été récemment proposé dans (Guan *et al.,* 2009) mais les auteurs se focalisent uniquement sur l'aspect personalisation de l'étiquetage (et non pas sur l'aspect annotation automatique).

Du point de vue des méthodes générales d'apprentissage, l'apprentissage dans les graphes est un champs très actif de recherche. Des méthodes ((Bilgic *et al.,* 2007)), (Maes *et al.,* 2009)) proposent des algorithmes itératifs pour l'étiquetage des noeuds d'un graphe. D'autres méthodes ((Zhou *et al.,* 2005), (Zhou *et al.,* 2003)) sont quant

à elles basées sur des extensions du problème de minimisation du risque empirique à travers l'ajout de termes de régularisation à une fonction objective classique. Cette dernière famille de modèle est celle dans laquelle s'inscrit notre travail. Tandis que toutes ces méthodes s'intéressent à la classification ou l'ordonnancement de noeuds d'un graphe, nous nous intéressons quant à nous à l'ordonnancement d'étiquettes pour chaque noeud (image) d'un graphe, ce qui est une tâche différente. Le travail le plus proche du notre est encore l'article (Abernethy *et al.,* 2008) qui s'intéresse à la détection de pages de spam dans un graphe de sites Web. Notre modèle peut être vu comme une extension de ce modèle dans le cas du ranking d'étiquettes (au lieu de la classification binaire).

6. Conclusion

Nous avons proposé un modèle qui est capable d'annoter automatiquement des images. Cette méthode prend en compte simultanément une information de contenu et une information relationnelle entre les images. Ces relations peuvent aussi bien être basées sur des similarités de contenu entre images, ou bien sur des relations explicites extraites du réseau social sous-jacent aux images traitées. Nous avons montré que notre modèle améliore nettement les performances d'un modèle de référence particulièrement quand il est utilisé avec des relations sociales comme la relation d'amitié par exemple. Cependant, ce modèle laisse quelques questions sans réponses : tout d'abord il n'est pas capable d'utiliser différents types de relations en même temps - tout comme les modèles existants par ailleurs. De plus, sa complexité est trop importante pour le traitement de très grandes masses d'images. Ces deux points sont aujourd'hui sous investigation.

7. Bibliographie

Abernethy J., Chapelle O., Castillo C., « Web spam identification through content and hyperlinks », *AIRWeb '08 : Proceedings of the 4th international workshop on Adversarial information retrieval on the web*, ACM, New York, NY, USA, 2008.

Barnard K., Duygulu P., Forsyth D., de Freitas N., Blei D. M., Jordan M. I., « Matching words and pictures », *J. Mach. Learn. Res.*, vol. 3, p. 1107-1135, 2003.

Bilgic M., Namata G., Getoor L., « Combining Collective Classification and Link Prediction », *ICDM Workshops*, p. 381-386, 2007.

Cao L., Luo J., Huang T. S., « Annotating photo collections by label propagation according to multiple similarity cues », *MM '08 : Proceeding of the 16th ACM international conference on Multimedia*, p. 121-130, 2008.

Chang E., Goh K., Sychay G., Wu G., « CBSA : Content-based Soft Annotation for Multimodal Image Retrieval Using Bayes Point Machines », *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 13, p. 26-38, 2003.

Frome A., Singer Y., Sha F., Jitendra M., « Learning Globally-Consistent Local Distance Functions for Shape-Based Image Retrieval and Classification », *Proceedings of IEEE 11th International Conference on Computer Vision*, p. 1-8, 2007.

Golder S. A., Huberman B. A., « Usage patterns of collaborative tagging systems », *J. Inf. Sci.*, vol. 32, n° 2, p. 198-208, 2006.

Guan Z., Bu J., Mei Q., Chen C., Wang C., « Personalized tag recommendation using graph-based ranking on multi-type interrelated objects », *SIGIR '09 : Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, ACM, New York, NY, USA, p. 540-547, 2009.

Har-Peled S., Roth D., Zimak D., « Constraint Classification : A New Approach to Multiclass Classification », *ALT*, p. 365-379, 2002.

Hironobu Y. M., Takahashi H., Oka R., « Image-to-Word Transformation Based on Dividing and Vector Quantizing Images With Words », *in Boltzmann machines, Neural Networks*, p. 405409, 1999.

Joachims T., « Optimizing search engines using clickthrough data », *KDD*, 2002.

Li J., Wang J. Z., « Real-Time Computerized Annotation of Pictures », *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, n° 6, p. 985-1002, 2008.

Maes F., Peters S., Denoyer L., Gallinari P., « Simulated Iterative Classification A New Learning Procedure for Graph Labeling », *ECML/PKDD (2)*, p. 47-62, 2009.

Matusiak K. K., « Towards user-centered indexing in digital image collections », *OCLC Systems & Services*, vol. 22, n° 4, p. 283-298, 2006.

Tong H., He J., Li M., Ma W.-Y., Zhang H.-J., Zhang C., « Manifold-ranking-based keyword propagation for image retrieval », *EURASIP J. Appl. Signal Process.*, 2006.

Zhou D., Bousquet O., Lal T. N., Weston J., Schölkopf B., « Learning with Local and Global Consistency », *NIPS*, 2003.

Zhou D., Huang J., Schölkopf B., « Learning from labeled and unlabeled data on a directed graph », *ICML*, p. 1036-1043, 2005.