
Apprentissage de métrique appliqué à la détection de changement de page Web et aux attributs relatifs

Marc T. Law* — **Nicolas Thome*** — **Stéphane Gançarski*** — **Matthieu Cord***

** Sorbonne Universités, UPMC Univ Paris 06, UMR 7606, LIP6, F-75005, Paris, France*

RÉSUMÉ. Nous proposons dans cet article un nouveau schéma d'apprentissage de métrique. Basé sur l'exploitation de contraintes qui impliquent des quadruplets d'images, notre approche vise à modéliser des relations sémantiques de similarités riches ou complexes. Nous étudions comment ce schéma peut être utilisé dans des contextes tels que la détection de régions importantes dans des pages Web ou la reconnaissance à partir d'attributs relatifs.

ABSTRACT. This paper introduces a novel distance metric learning framework. Working with inequality constraints involving quadruplets of images, our approach aims at efficiently modeling similarity for rich or complex semantic label relationships. We study how our metric learning scheme can be exploited in contexts such as detection of important regions in Webpages or recognition with relative attributes.

MOTS-CLÉS : apprentissage de métrique, reconnaissance d'image, détection de changement

KEYWORDS: distance metric learning, image recognition, webpage change detection

1. Introduction

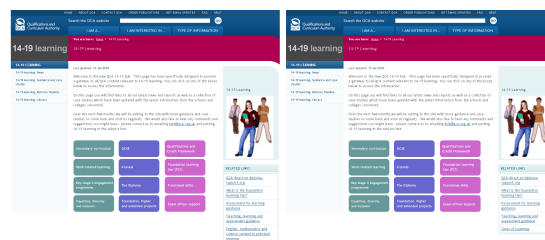
La préservation du patrimoine culturel est une préoccupation majeure de nombreuses civilisations. Au-delà du simple rassemblement d'objets culturels, les méthodes de conservation et de diffusion utilisées sont problématiques. Aujourd'hui, notre culture crée et diffuse de plus en plus d'information sous forme numérique, et le plus grand répertoire actuel de contenu (journaux d'informations, articles, documents publiés, etc...) est le Web. Des organismes, tels qu'*Internet Archive*¹, cherchent à préserver ces objets numériques et les rendre disponibles à des chercheurs, historiens, universitaires de notre génération, et de celles à venir. *Internet Archive* recense 240 milliards de pages Web archivées de 1996 à 2012. Avec l'explosion du volume d'information disponible, le contenu d'Internet change et croît de façon permanente. Le rythme actuel de changement du Web est tel que l'espérance de vie du contenu d'une page Web est de 77 jours. Si aucun effort n'est fourni pour préserver ce contenu, il sera alors entièrement perdu et irrécupérable.

Les deux problématiques majeures de l'organisme *Internet Archive* sont : (1) comment préserver et diffuser les données qui ont déjà été capturées et (2) comment stocker l'énorme volume de données du Web. Ceci concerne d'une part les méthodes d'indexation et d'encodage physique des données, et d'autre part les choix pertinents qui doivent être faits par rapport à la croissance exponentielle du Web. Face aux contraintes techniques qui empêchent les archivistes du Web de capturer les contenus d'information à tout moment, des études de comportement de sites Web (Adar *et al.*, 2009) se sont imposées pour déterminer des stratégies intelligentes de visite de robots d'indexation. Parmi ces études de comportement, la méthode qui s'est montrée la plus robuste consiste à faire plusieurs comparaisons de versions successives d'une même page Web et identifier à quelle fréquence un changement sémantique est apparu.

Nous illustrons la notion de changement sémantique dans le contexte de l'archivage à travers deux exemples. La Figure 1 présente deux versions successives d'une même page Web dont le seul changement concerne des hyperliens aléatoires qui peuvent être ignorés. Puisque l'information principale de la page est identique entre les deux versions, celles-ci peuvent être considérées similaires. Un robot d'indexation ne nécessite alors pas de les visiter toutes deux. La Figure 2 présente en revanche deux versions successives dont l'information partagée par la page (qui concerne l'actualité) est différente. Ces deux versions sont donc considérées dissimilaires et doivent toutes deux être sauvegardées.

Dans cette analyse de ressemblance sémantique entre données multimedia, les méthodes de comparaison sont limitées par le type d'information qu'elles exploitent. En effet, une page Web peut être comparée en se limitant à son code source, ou en utilisant aussi les différentes ressources (images, feuilles de style etc...) qu'elle charge et qui représentent de l'information. Très tôt (Cai *et al.*, 2003 ; Song *et al.*, 2004), le

1. <http://archive.org/index.php>



Zoom sur l'unique différence entre les deux versions :

RELATED LINKS	RELATED LINKS
QCA direct on diploma-support.org	QCA direct on diploma-support.org
What is the foundation learning tier?	What is the foundation learning tier?
Assessment for learning guidance	Assessment for learning guidance
Teaching, learning and assessment guidance	Teaching, learning and assessment guidance
English, mathematics and science content in principal learning	Lines of Learning

Figure 1. Un exemple de deux versions similaires successives d'une page dont la différence n'est pas assez importante pour nécessiter une réindexation.

rendu visuel de la page a été exploité, que ce soit pour la segmentation de page (Cai *et al.*, 2003), ou la localisation de contenu important (Song *et al.*, 2004). Il permet de prendre en compte les similitudes entre différentes régions d'une même page, et a donc de même été utilisé pour comparer des versions de page dans une finalité d'archivage (Pehlivan *et al.*, 2010 ; Ben Saad et Gançarski, 2011). Toutefois, ces méthodes utilisent des mesures de similarité prédéfinies ou dont les paramètres sont déterminés de façon *ad hoc*. Or une famille de méthodes de l'apprentissage automatique vise à apprendre une similarité ou métrique (Xing *et al.*, 2002 ; Weinberger et Saul, 2009) entre paires d'éléments. Cet article vise à fournir des mesures de similarité "sémantique" en se concentrant sur ce contexte d'apprentissage de similarité. Nous cherchons à tirer profit du très grand nombre de données non annotées disponibles à l'aide de méthodes d'apprentissage. Notamment, nous proposons une nouvelle approche qui exploite les relations temporelles entre les différentes versions pour extraire de façon totalement automatique les régions de pertinence d'une page Web.

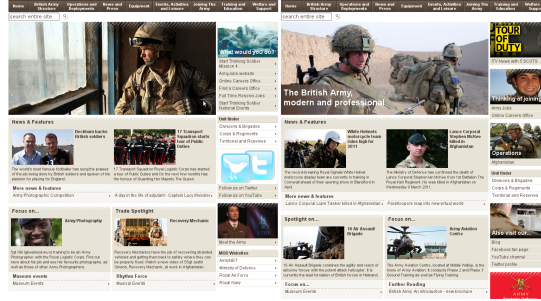


Figure 2. Deux versions de page Web dissimilaires qui contiennent des informations différentes et nécessitent donc une réindexation.

2. Apprentissage de métrique

L'apprentissage de similarité ou de distance consiste à apprendre une mesure de distance sur un espace de données d'entrée. Il s'intéresse souvent à préserver la relation de distance au sein de la base d'apprentissage entre paires d'exemples similaires/dissimilaires. L'apprentissage de métrique pour la classification et la recherche d'image s'effectue principalement selon deux familles de méthodes :

- Les méthodes par paire d'exemples (Xing *et al.*, 2002 ; Mignon et Jurie, 2012) : elles travaillent généralement sur deux ensembles, appelés \mathcal{S} et \mathcal{D} , de paires d'objets similaires et dissimilaires, respectivement. Le but est de minimiser la distance entre paires d'exemples similaires (dans \mathcal{S}) tout en séparant les exemples dissimilaires (dans \mathcal{D}). Elle se traduit souvent par apprendre une distance D et un seuil γ tels que $D(p_i, p_j) < \gamma$ si $(p_i, p_j) \in \mathcal{S}$ et $D(p_i, p_j) > \gamma$ si $(p_i, p_j) \in \mathcal{D}$.
- Les méthodes par triplet (p_i, p_i^+, p_i^-) : ces méthodes (Frome *et al.*, 2006 ; Chechik *et al.*, 2010 ; Weinberger et Saul, 2009) dérivent des méthodes par paire, elles s'intéressent à des paires d'exemples $(p_i, p_i^+) \in \mathcal{S}$ et $(p_i, p_i^-) \in \mathcal{D}$. Le but est d'apprendre une distance D qui respecte $D(p_i, p_i^+) < D(p_i, p_i^-)$.

Les deux modèles de dissimilarité les plus populaires en apprentissage de métrique sont la distance de Mahalanobis, et une combinaison linéaire de distances élémentaires. Dans le cas de la combinaison linéaire de distances, il s'agit d'apprendre

une distance $D_{\mathbf{w}}$ sous la forme $D_{\mathbf{w}}(p_i, p_j) = \mathbf{w}^T \mathbf{d}(p_i, p_j) = \sum_{k=1}^N w_k d_k(p_i, p_j)$ où

d_1, \dots, d_N sont des distances élémentaires entre les images p_i et p_j et $\mathbf{w} \in \mathbb{R}^N$ est leur vecteur de pondération. Dans (Frome *et al.*, 2006), la contrainte $\mathbf{w} \geq \mathbf{0}$ (les éléments de \mathbf{w} sont non négatifs) est forcée pour respecter $D_{\mathbf{w}}$ non négative.

Une distance de type Mahalanobis $D_{\mathbf{W}}$ s'écrit sous la forme :

$$D_{\mathbf{W}}^2(p_i, p_j) = \Phi(p_i, p_j)^T \mathbf{W} \Phi(p_i, p_j) \quad [1]$$

où $\mathbf{W} \in \mathbb{R}^{d \times d}$ est une matrice symétrique semi-définie positive (SDP), et $\Phi(p_i, p_j) \in \mathbb{R}^d$ est un vecteur de dissimilarité entre p_i et p_j . On choisit généralement $\Phi(p_i, p_j) = \mathbf{x}_i - \mathbf{x}_j$ où \mathbf{x}_i et \mathbf{x}_j sont les représentations vectorielles des images p_i et p_j . A l'origine, le terme de distance de Mahalanobis était utilisé pour décrire les formes quadratiques de distributions gaussiennes où la matrice \mathbf{W} joue le rôle de l'inverse de la matrice de covariance. Dans cet article, nous autorisons toute matrice semi-définie positive.

Une matrice symétrique \mathbf{W} est SDP ssi toutes ses valeurs propres sont non négatives. Elle peut aussi être décomposée en $\mathbf{W} = \mathbf{L}^T \mathbf{L}$ où $\mathbf{L} \in \mathbb{R}^{e \times d}$ où e dépend du rang de \mathbf{W} . En particulier, $e \geq \text{rang}(\mathbf{L}) = \text{rang}(\mathbf{W})$. On peut alors réécrire :

$$D_{\mathbf{W}}^2(p_i, p_j) = \Phi(p_i, p_j)^T \mathbf{L}^T \mathbf{L} \Phi(p_i, p_j) = \|\mathbf{L} \Phi(p_i, p_j)\|_2^2 \quad [2]$$

On remarque qu'à partir de toute transformation linéaire paramétrée par \mathbf{L} , on peut induire une distance de Mahalanobis. Réciproquement, à partir de n'importe quelle distance de Mahalanobis, on peut déduire une transformation linéaire paramétrée par \mathbf{L} (même si la décomposition de \mathbf{W} n'est en général pas unique). Si $\Phi(p_i, p_j) = \mathbf{x}_i - \mathbf{x}_j$, calculer la distance de Mahalanobis dans l'espace d'entrée correspond alors à calculer la distance euclidienne classique entre \mathbf{x}_i et \mathbf{x}_j dans l'espace induit par \mathbf{L} .

Weinberger et Saul (Weinberger et Saul, 2009) parlent d'équivalence entre l'apprentissage de distance de Mahalanobis et l'apprentissage de transformation linéaire. Apprendre une transformation linéaire (*e.g.* avec une analyse en composantes principales) correspond à de l'apprentissage de métrique. Pour éviter d'apprendre des modèles trop complexes, certains (Mignon et Jurie, 2012 ; Weinberger et Saul, 2009) optimisent l'apprentissage par rapport à $\mathbf{L} \in \mathbb{R}^{n \times d}$ où $n \ll d$ pour avoir le rang de \mathbf{W} inférieur ou égal à n . Le nombre de paramètres à apprendre passe aussi de $O(d^2)$ à $O(nd) \simeq O(d)$. Cependant, optimiser par rapport à \mathbf{L} conduit généralement à des problèmes d'optimisation non convexes. Optimiser par rapport à \mathbf{W} mène vers des problèmes convexes, mais il faut garantir à chaque itération que \mathbf{W} a ses valeurs propres non négatives, ce qui requiert une décomposition spectrale computationnellement lourde.

Dans cet article, nous proposons d'apprendre une métrique à partir de contraintes sur des quadruplets d'images. Ces contraintes sont en fait une généralisation des contraintes classiques par paire ou triplet (cf Section 4.1 dans (Law *et al.*, 2014)). Nous illustrerons quelques exemples applicatifs où de telles contraintes sont plus adaptées que des contraintes sur des paires ou triplets d'images.

3. Apprentissage de distance basé sur les quadruplets d'images

Nous présentons dans cette section une méthode qui permet de comparer de façon efficace des quadruplets d'images. Nous prouvons expérimentalement le bon fonctionnement de la méthode sur deux applications : (1) la détection de régions de changement important dans une page Web sans aucune annotation humaine, (2) la reconnaissance de visage et de scène à l'aide d'attributs relatifs.

3.1. Présentation du problème

Attribuer un label "similaire" ou "dissimilaire" entre deux objets est parfois difficile, il est souvent plus facile de dire "les images p_i et p_j sont plus similaires que ne le sont p_k et p_l ". Ce problème survient dans différents contextes, comme par exemple l'étude de modèles graphiques visant à uniformiser la perception des différences de couleurs (Perrot *et al.*, 2014) : "telles couleurs sont aussi similaires que telles autres". Nous présentons donc le modèle **Qwise** (*quadruplet-wise*) que nous avons proposé dans (Law *et al.*, 2013). Notre modèle prend en compte des quadruplets d'images pour minimiser une certaine fonction d'optimisation décrite dans la suite de cet article.

Soient \mathcal{P} un ensemble d'images et une fonction de dissimilarité cible $D : \mathcal{P} \times \mathcal{P} \rightarrow \mathbb{R}$ que nous cherchons à approcher, en notant $D(p_i, p_j) = D_{ij}$, nous définissons les deux ensembles de quadruplets \mathcal{A} et \mathcal{B} , selon les contraintes :

$$\forall (p_i, p_j, p_k, p_l) \in \mathcal{A}, D_{ij} > D_{kl} \quad [3]$$

$$\forall (p_i, p_j, p_k, p_l) \in \mathcal{B}, D_{ij} \geq D_{kl} \quad [4]$$

L'équation 3 traduit une inégalité stricte entre les dissimilarités de (p_i, p_j) et (p_k, p_l) , l'équation 4 traduit une inégalité non stricte. Notons que $D_{ij} = D_{kl}$ se réécrit sous les deux contraintes $D_{ij} \leq D_{kl}$ et $D_{ij} \geq D_{kl}$. En cherchant à approximer D par une distance de Mahalanobis formulée comme dans l'Equation 1 par $D_{\mathbf{W}}^2$, notre problème revient à vouloir maximiser le nombre de contraintes suivantes respectées :

$$\forall (p_i, p_j, p_k, p_l) \in \mathcal{A}, D_{\mathbf{W}}^2(p_i, p_j) - D_{\mathbf{W}}^2(p_k, p_l) \geq 1 \quad [5]$$

$$\forall (p_i, p_j, p_k, p_l) \in \mathcal{B}, D_{\mathbf{W}}^2(p_i, p_j) - D_{\mathbf{W}}^2(p_k, p_l) \geq 0 \quad [6]$$

Il est à noter que le choix de la marge (1 dans l'Equation 5) peut être déterminé par validation croisée sur les données considérées.

Approximer la solution en intégrant des fonctions de coût de type *hinge loss* sur $\mathcal{A} \cup \mathcal{B}$ est simple mais (1) est généralement computationnellement coûteux si l'on dérive par rapport à \mathbf{W} , car il faut reprojeter la solution sur le cône des matrices semi-définies positives à chaque itération. (2) En décomposant $\mathbf{W} = \mathbf{L}^T \mathbf{L}$, faire une descente de gradient par rapport à \mathbf{L} mène vers des problèmes non convexes.

3.2. Distance de Mahalanobis avec matrice diagonale

Pour rendre le problème à la fois convexe et efficace, une solution est de contraindre la matrice \mathbf{W} à être diagonale. Assurer que \mathbf{W} est SDP est alors équivalent à forcer \mathbf{W} à avoir ses éléments (diagonaux) non négatifs car les éléments de sa diagonale sont aussi ses valeurs propres dans ce cas. Nous notons $\mathbf{w} = \text{Diag}(\mathbf{W})$, dans le cas où \mathbf{W} est diagonale, nous avons l'égalité :

$$D_{\mathbf{W}}^2(p_i, p_j) = \Phi(p_i, p_j)^T \mathbf{W} \Phi(p_i, p_j) = \mathbf{w}^T [\Phi(p_i, p_j) \circ \Phi(p_i, p_j)] = \mathfrak{D}_{\mathbf{w}}(p_i, p_j) \quad [7]$$

où \circ est le produit matriciel de Hadamard (produit terme à terme de matrices). Nous notons $\Psi(p_i, p_j) = \Phi(p_i, p_j) \circ \Phi(p_i, p_j)$ dans ce cas.

En restreignant ainsi \mathbf{W} , et en notant $q = (p_i, p_j, p_k, p_l)$ et $\mathbf{z}_q = \mathbf{z}_{ijkl} = \Psi(p_i, p_j) - \Psi(p_k, p_l)$, nous avons $\forall q \in \mathcal{P} \times \mathcal{P} \times \mathcal{P} \times \mathcal{P}, \mathbf{w}^T \mathbf{z}_q = \mathfrak{D}_{\mathbf{w}}(p_i, p_j) - \mathfrak{D}_{\mathbf{w}}(p_k, p_l)$. Les équations 5 et 6 peuvent alors se réécrire $\forall q \in \mathcal{A}, \mathbf{w}^T \mathbf{z}_q \geq 1$ et $\forall q \in \mathcal{B}, \mathbf{w}^T \mathbf{z}_q \geq 0$. Nous définissons des fonctions de coût L_ϵ adaptées à chaque ensemble de quadruplets \mathcal{A} ou \mathcal{B} selon la valeur de marge de sécurité ϵ utilisée. Nous pouvons par exemple définir $L_\epsilon(t) = \max(0, \epsilon - t)$ où L_1 correspond à la fonction *hinge loss*. Notre problème d'optimisation final, adapté du ranking SVM (Joachims, 2002), vise à minimiser $\mathbf{w}^T \mathbf{w} = \|\mathbf{w}\|_2^2$ et la fonction de coût globale associée aux ensembles \mathcal{A} et \mathcal{B} . Il s'écrit sous la forme :

$$\min_{\mathbf{w}} \frac{1}{2} \|\mathbf{w}\|_2^2 + C \left(\sum_{q \in \mathcal{A}} L_1(\mathbf{w}^T \mathbf{z}_q) + \sum_{q \in \mathcal{B}} L_0(\mathbf{w}^T \mathbf{z}_q) \right) \quad [8]$$

Le problème est linéaire par rapport au nombre de contraintes (cardinal de $\mathcal{A} \cup \mathcal{B}$). Nous résolvons ce problème convexe par rapport à \mathbf{w} dans le primal par une descente de gradient. Aussi, pour forcer \mathbf{W} à être SDP, nous forçons la contrainte $\mathbf{w} \geq \mathbf{0}$. A noter que le problème décrit dans l'Equation 8 peut être étendu aux noyaux (Chapelle, 2007).

4. Détection automatique de régions importantes dans les pages Web

Nous proposons de reconnaître les régions où se produisent les changements importants dans une page Web en n'ayant besoin d'aucune annotation humaine. Pour cela, nous supposons une monotonie des changements. Nous représentons les paires de pages par un ensemble de distances locales. Nous segmentons chaque page en plusieurs blocs, et nous calculons les distances locales entre ces blocs. Pour faciliter la segmentation, nous effectuons un découpage régulier de l'image par une grille de 8×8 ou 10×10 blocs comme illustré dans la Figure 4 (a). Les distances locales sont alors calculées entre blocs (de deux versions) qui se superposent.

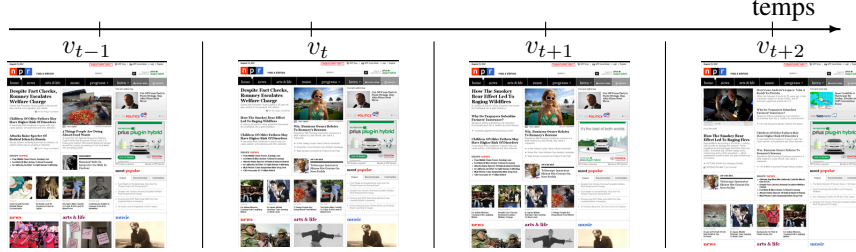


Figure 3. Versions successives de la page d'accueil de National Public Radio (NPR).

4.1. Comparaison des distances entre versions selon leurs relations temporelles

Nous notons v_t une version capturée à l'instant t . Dans la figure 3, les versions $(v_{t-1}, v_t, v_{t+1}, v_{t+2})$ sont quatre versions successives d'une même page Web, capturées à 1 heure d'intervalle. Sans même s'intéresser à annoter si les versions successives (v_{t-1}, v_t) , (v_t, v_{t+1}) ou (v_{t+1}, v_{t+2}) sont similaires ou dissimilaires, nous observons que (v_t, v_{t+1}) sont plus similaires que ne le sont (v_t, v_{t+2}) , voire (v_{t-1}, v_{t+2}) .

Soit D une dissimilarité, nous supposons que la plupart des contraintes suivantes sont respectées pour des versions successives d'une même page :

$$\forall r \leq t, \forall s \geq t+1, D(v_t, v_{t+1}) \leq D(v_r, v_s) \quad [9]$$

Toutefois, puisque nous cherchons à approximer D par une pseudo-métrique D_W où W est une matrice semi-définie positive (et diagonale), la méthode la plus simple de respecter ces contraintes consiste à choisir W avec tous ses éléments nuls. Pour éviter ce cas dégénéré, nous supposons qu'il existe un paramètre $\gamma > 0$ tel que $\forall r(v_r, v_{r+\gamma}) \in \mathcal{D}$, à partir duquel nous avons l'inégalité stricte :

$$\forall r \leq t, \forall r + \gamma \geq t+1, D(v_t, v_{t+1}) < D(v_r, v_{r+\gamma}) \quad [10]$$

Que la paire (v_t, v_{t+1}) soit similaire ou dissimilaire, sa dissimilarité est strictement plus faible que celle de $(v_r, v_{r+\gamma})$. La valeur optimale de γ dépend évidemment de la page en question, Adar *et al.* (Adar *et al.*, 2009) observent par exemple que les sites gouvernementaux (dont l'URL se termine par ".gov") ont des changements moins fréquents que ceux ayant une URL se terminant par ".com", ou encore que les sites souvent revisités (populaires) ont tendance à souvent changer. Dans ce dernier cas, cela peut s'expliquer par le fait que les humains agissent comme des *crawlers* intelligents. Ces informations de fréquentation peuvent s'obtenir par des services comme Google Analytics, et permettent ainsi de déterminer heuristiquement la valeur du γ . Bien entendu, avoir un γ très grand n'est pas si problématique car la contrainte de l'équation 9 est respectée si l'équation 10 est respectée.

Intuition de la méthode : Nous représentons notre distance comme une somme pondérée des distances entre régions superposées de deux versions différentes. Les

contraintes du type $D(v_t, v_{t+1}) \leq D(v_r, v_s)$ sont censées ignorer les contenus aléatoires et périodiques, comme les publicités. Nous l’observons dans la Figure 3 : la publicité de voiture est différente dans la version v_{t+1} mais redevient identique dans v_{t+2} , les régions dans lesquelles elle apparaît violent donc les contraintes $D(v_t, v_{t+1}) \leq D(v_t, v_{t+2})$ et $D(v_t, v_{t+1}) \leq D(v_{t-1}, v_{t+2})$.

Les distance $D(v_t, v_{t+1}) < D(v_r, v_{r+\gamma})$ pénalisent quant à elles les menus et autres contenus qui changent très rarement, voire jamais. En effet, un changement est attendu entre les versions v_r et $v_{r+\gamma}$, les menus qui changent très rarement violent donc ces contraintes et sont ignorés.

4.2. Résultats expérimentaux

Ensemble de données : Pour évaluer notre approche, nous avons crawlé les pages d’accueil² de CNN, BBC, NPR et le New York Times pendant environ 50 jours. Le crawling est effectué chaque heure : v_{t+1} est visité 1 heure après v_t , comme c’est généralement fait en Web crawling (Adar *et al.*, 2009 ; Ben Saad et Gançarski, 2011). Pour une évaluation quantitative, nous avons manuellement annoté les paires successives de versions (v_t, v_{t+1}) :

- Dissimilaires : l’information principale de la page (une *news* importante) a changé.
- Identiques : il n’est apparu aucun changement.
- Similaires : le changement effectué n’est pas celui de la *news* principale (changement de publicité).
- Ambiguës : l’annotation est difficile.

Nous nous sommes restreints à des sites d’actualité pour faciliter l’annotation. Les pages identiques et ambiguës ne sont pas prises en compte lors de l’évaluation.

Mesure d’évaluation : Une bonne distance doit avoir une plus faible valeur pour les éléments similaires que pour les éléments dissimilaires. Pour mesurer la qualité de la distance, nous utilisons l’*Average Precision* (AP) des classes Similaires et Dissimilaires :

- AP_S : nous rangeons les distances des couples test (v_t, v_{t+1}) dans l’ordre croissant et calculons l’AP pour la classe des similaires.
- AP_D : nous rangeons les distances des couples test (v_t, v_{t+1}) dans l’ordre décroissant et calculons l’AP pour la classe des dissimilaires.

Le *Mean Average Precision* (MAP) est en fait la moyenne de AP_S et AP_D .

² <http://www.cnn.com>, <http://www.bbc.co.uk>, <http://www.npr.org>, <http://www.nytimes.com>

Site web	CNN	NPR	New York Times	BBC
MAP de la distance euclidienne	77.0 ±0.5	92.9 ±0.3	74.6 ±0.5	83.9 ±0.4
MAP de LMNN	85.2 ±1.8	95.2 ±0.9	86.1 ±2.0	86.3 ±0.6
MAP de Qwise	88.6 ±2.9	96.5 ±0.4	88.9 ±4.6	86.1 ±0.8

Tableau 1. Résultats d’Average Precision (en %) sur la base de test sur les pages d’accueil de CNN, NPR, New York Times et BBC.

Nous découpons la base en 10 splits différents formés de 5 jours successifs pour l’apprentissage, et des 45 jours restants pour le test. Il n’y a aucune annotation humaine lors de l’apprentissage, seules les relations temporelles décrites plus haut sont utilisées.

Descripteurs visuels : Nous considérons les captures d’écran des pages comme des images. Seule la partie visible des pages est prise en compte puisqu’elle contient généralement l’information principale de la page (Song *et al.*, 2004). Nous utilisons des descripteurs GIST (Oliva et Torralba, 2001) qui segmentent l’image par une grille de taille $m \times m$. Nous formulons $\Psi(v_r, v_s) \in \mathbb{R}^{m^2}$ comme un vecteur dont chaque élément correspond à la distance euclidienne au carré des orientations du GIST qui tombent dans la même région de l’image.

Paramètres : Comme paramètre de segmentation, nous choisissons $m = 10$ pour les résultats quantitatifs de Tableau 1. Pour générer nos contraintes, nous extrayons des quadruplets décrits dans les équations 9 et 10 tels que $r \geq t - 6$, $s \leq t + 7$, $\gamma = 4$.

Baselines : Nous comparons notre modèle à la distance euclidienne ($\mathbf{W} = I_d$ dans l’équation 1) obtenue sans apprentissage, et à LMNN (Weinberger et Saul, 2009) qui est en fait appris sur le sous-ensemble formé de triplets, des quadruplets que nous avons extraits de l’Equation 10. Le modèle Qwise comprend donc plus de contraintes que LMNN.

Les résultats quantitatifs sont rapportés dans le Tableau 1, notre méthode surpasse les deux autres méthodes sur la plupart des sites. LMNN, appris sur moins de contraintes obtient des résultats faiblement moins bons sauf sur BBC qui a une stratégie d’affichage assez particulière. En effet, une bannière "*breaking news*" apparaît lors de l’apparition de news importantes (dans la seule région où apparaissent les changements importants), et disparaît quelque temps plus tard. Ceci crée chez nous de fausses détections car le contenu textuel est identique. Une solution à ce problème serait d’intégrer les similarités textuelles pour constater que le texte est identique. L’information visuelle seule n’est pas capable d’analyser si ces changements sont importants ou non sans annotation. Toutefois, elle est bien capable de détecter les zones dans lesquels les changements importants se produisent comme décrit ci-dessous.

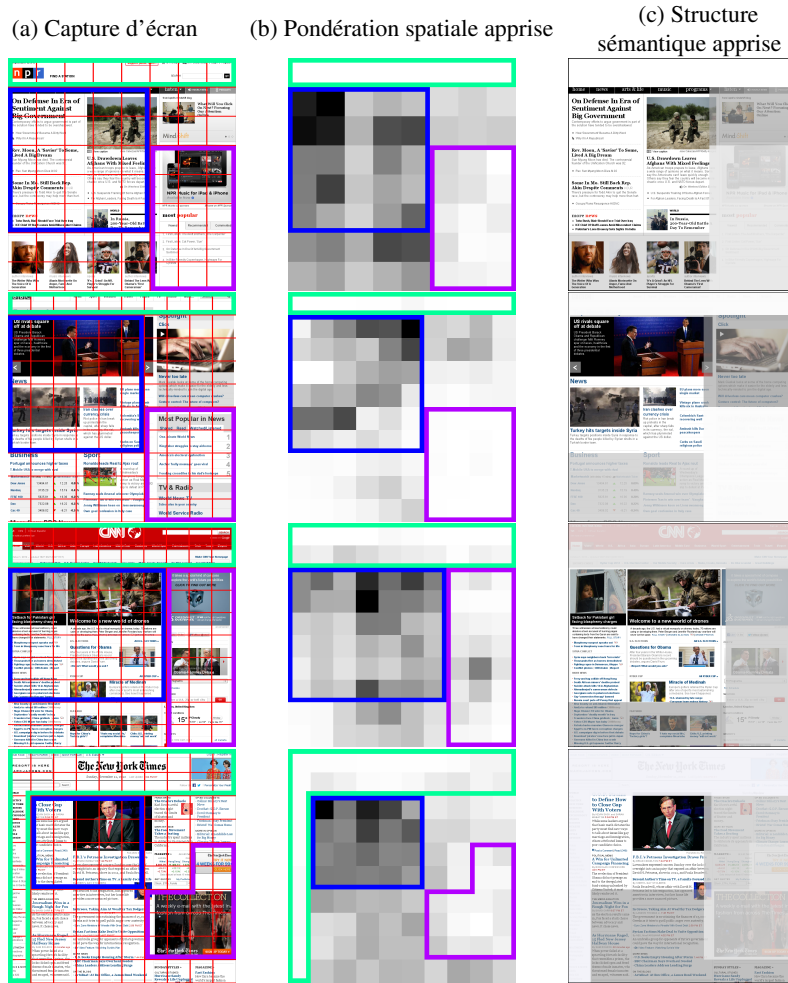


Figure 4. Cartes de dynamique des pages d'accueil de NPR, BBC, CNN et New York Times. (Gauche) Capture d'écran avec les parties importantes (actualité principale) en bleu, les parties à ignorer en vert (bannière et menu) et en violet (contenu aléatoire, publicité). (Milieu) Pondération spatiale de la dynamique des pages (les valeurs les plus grandes sont foncées). (Droite) Visualisation de la structure sémantique et spatiale apprise. De la transparence est ajoutée sur les régions selon leur pondération moyenne apprise.

Visualisation : Puisque nous avons appris un vecteur $\mathbf{w} \geq \mathbf{0}$ qui pondère chaque région de l'image, nous sommes capables de visualiser l'importance apprise pour chaque région. En effet, les éléments de \mathbf{w} les plus proches de 0 correspondent aux régions ignorées de la page. Nous affichons la pondération spatiale apprise dans la figure 4 (b), où les régions les plus importantes sont sombres. Nous ajoutons sur chaque capture une région :

- bleue : l'information principale de la page.
- verte : la bannière et le menu
- violette : du contenu aléatoire, par exemple de la publicité.

Nous vérifions sur les figures que les menus, bannières et publicités sont ignorés et que l'information principale de la page correspond à la zone la plus foncée. Nous rappelons que tout cela est appris sans aucune annotation humaine.

Des extensions de notre modèle qui se baseraient sur une combinaison des informations structurelle et visuelle de la page peuvent être faites comme dans (Law *et al.*, 2012).

5. Apprentissage de descripteurs sémantiques

Le but de cette section est de faire de la classification d'image à partir d'information haut-niveau sur les classes. Ces informations haut-niveau sont appelées "attributs" et servent dans notre contexte à représenter les images dans un nouvel espace de description.

5.1. *Attributs relatifs*

Nous présentons ici notre modèle Qwise appliqué au contexte des attributs relatifs (Parikh et Grauman, 2011). Les attributs sont des concepts haut-niveau (concepts sémantiques) pour décrire des classes. Par exemple, sur la base *Animal With Attributes* (AWA) (Lampert *et al.*, 2009), des valeurs binaires sont attribuées à chaque classe d'animaux pour savoir si ses individus "mangent du plancton", "vivent dans l'océan", "sont chasseurs", "sont jaunes"...

Dans (Parikh et Grauman, 2011), les annotations d'attributs portent sur les classes et non les images, ce qui requiert beaucoup moins d'annotations manuelles des données. A chaque image est alors attribuée l'annotation affectée à sa classe. Plutôt qu'utiliser des valeurs binaires d'annotation d'attributs, Parikh et Grauman (Parikh et Grauman, 2011) proposent d'annoter les attributs par comparaison de classes. Par exemple dans la figure 5, elles utilisent des annotations du type $(e) \prec (f) \sim (g) \prec (h)$ où $(e) \prec (f)$ signifie que les images de la classe (f) ont une présence plus importante de l'attribut "présence de sourire" que les images de la classe (e) , et $(f) \sim (g)$ signifie que les deux classes (f) et (g) ont une présence équivalente de l'attribut "présence de sourire". Pour chaque attribut, Parikh et Grauman ne considèrent que des relations

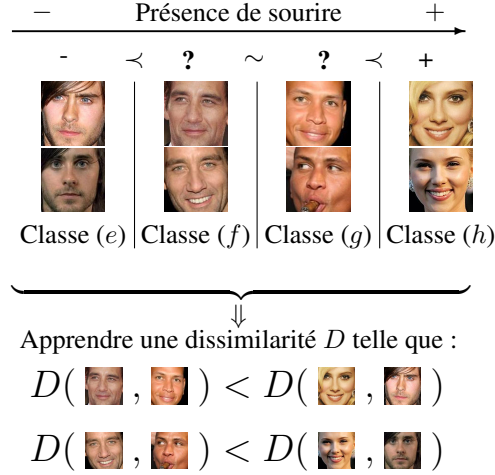


Figure 5. Stratégie *Qwise* sur 4 classes de visage rangées selon le degré de présence de sourire. Au lieu de travailler sur des relations par paire qui présentent des défauts, la stratégie *Qwise* définit des contraintes sur des ensembles de 4 images pour exprimer que les dissimilarités entre exemples des classes (f) et (g) doivent être plus petites que les dissimilarités entre exemples de (e) et (h).

entre deux classes. Or il est évident que les images des classes (f) et (g) ne sourient pas toujours autant (cf 2ème ligne de la Figure 5). Plutôt que de forcer une équivalence entre les classes (f) et (g), nous considérons que l'écart de sourire entre les classes (e) et (h) est plus important que l'écart entre (f) et (g).

En prenant en compte une dissimilarité signée pour chaque attribut a_m , elles s'intéressent à respecter les contraintes $D_m(p_i, p_j) > 0$ si p_i est plus souriant que p_j , ou $D_m(p_i, p_j) = 0$ si p_i est aussi souriant que p_j . Elles proposent pour cela d'approximer D_m par une dissimilarité $\mathcal{D}_{\mathbf{w}_m}(p_i, p_j) = \mathbf{w}_m^\top (\mathbf{x}_i - \mathbf{x}_j)$ où $\mathbf{x}_i \in \mathbb{R}^d$ est la représentation vectorielle bas-niveau de p_i et $\mathbf{w}_i \in \mathbb{R}^d$ un vecteur de pondération. Nous proposons, à la place, de considérer des contraintes de la forme $D_m(p_i, p_j) > D_m(p_k, p_l)$. Par exemple, dans l'exemple de la figure 5, nous aurions $(p_i, p_j, p_k, p_l) \in (h) \times (e) \times (g) \times (f)$.

Soit M le nombre d'attributs, elles proposent de représenter chaque image p_i par un vecteur de représentation haut-niveau $\mathbf{h}_i = [\mathbf{w}_1^\top \mathbf{x}_i, \dots, \mathbf{w}_m^\top \mathbf{x}_i, \dots, \mathbf{w}_M^\top \mathbf{x}_i]^\top$ dont chaque élément $\mathbf{w}_m^\top \mathbf{x}_i$ représente le degré de présence de l'attribut a_m dans l'image p_i . Leur formalisme correspond à apprendre une transformation linéaire paramétrée par $\mathbf{L} \in \mathbb{R}^{M \times d}$ telle que la $m^{\text{ième}}$ ligne de \mathbf{L} soit \mathbf{w}_m^\top . On a alors $\mathbf{h}_i = \mathbf{L}\mathbf{x}_i$.

Attributs OSR	Ordre relatif sur les classes
Naturel	$T \prec I \sim S \prec H \prec C \sim O \sim M \sim F$
Ouvert	$T \prec F \prec I \sim S \prec M \prec H \sim C \sim O$
Perspective	$O \prec C \prec M \sim F \prec H \prec I \prec S \prec T$
Grands objets	$F \prec O \prec M \prec I \sim S \prec H \sim C \prec T$
Plan diagonal	$F \prec O \prec M \prec C \prec I \sim S \prec H \prec T$
Plan de proximité	$C \prec M \prec O \prec T \sim I \sim S \sim H \sim F$

Tableau 2. Ordres relatifs (annotés sur les classes) pour la bases OSR (coast (C), forest (F), highway (H), inside-city (I), mountain (M), open-country (O), street (S) and tall-building (T)).

5.2. Résultats expérimentaux

Nous nous comparons dans une tâche de classification en utilisant le même setup et sur les mêmes ensembles de données que dans (Parikh et Grauman, 2011).

Ensembles de données : Nous faisons nos expériences sur les ensembles de données : **Outdoor Scene Recognition (OSR)** (Oliva et Torralba, 2001) qui contient 2688 images de 8 catégories de scènes et un sous-ensemble de **Public Figure Face (Pub-Fig)** (Kumar *et al.*, 2009) qui contient 771 images de 8 catégories de visages. Nous utilisons les descripteurs visuels rendus disponibles par (Parikh et Grauman, 2011) : un descripteur GIST (Oliva et Torralba, 2001) de 512 dimensions pour OSR et une concaténation du descripteur GIST et un histogramme de couleur de 45 dimensions pour PubFig. Nous utilisons, en plus des labels d’appartenance aux classes, les annotations d’attributs relatifs (cf Tableau 2).

Baselines : Nous nous comparons aux deux baselines suivantes (1) le problème défini dans Parikh et Grauman (Parikh et Grauman, 2011) et décrit dans la Section 5, (2) l’algorithme **Large Margin Nearest Neighbor (LMNN)** (Weinberger et Saul, 2009), une méthode de référence de *metric learning*. A chaque image est assigné un ensemble de k plus proches voisins cibles, l’algorithme fait en sorte d’apprendre une transformation linéaire telle que la distance euclidienne de chaque image à ses k -ppv dans l’espace induit soit plus petite que celle des images des autres classes.

Méthodes proposées : Nous proposons trois méthodes :

- La méthode appelée Qwise qui consiste à étendre les contraintes par paire de Parikh et Grauman (Parikh et Grauman, 2011) à des contraintes avec quadruplets. Nous nous comparons à nombre de contraintes égal, pour chacune de leurs contraintes par paire, nous créons une contrainte par quadruplet.
- Une méthode appelée Qwise + LMNN : l’espace appris avec la méthode Qwise est utilisée comme espace d’entrée de LMNN qui utilise uniquement l’information de classe pour dissocier les images.

	OSR	Pubfig
Code de Parikh (Parikh et Grauman, 2011)	$71.3 \pm 1.9\%$	$71.3 \pm 2.0\%$
LMNN-G	$70.7 \pm 1.9\%$	$69.9 \pm 2.0\%$
LMNN	$71.2 \pm 2.0\%$	$71.5 \pm 1.6\%$
Qwise	$74.1 \pm 2.1\%$	$74.5 \pm 1.3\%$
Qwise + LMNN-G	$74.6 \pm 1.7\%$	$76.5 \pm 1.2\%$
Qwise + LMNN	$74.3 \pm 1.9\%$	$77.6 \pm 2.0\%$
Qwise + LMNN + Fantope	$75.0 \pm 2.0\%$	$77.6 \pm 1.1\%$

Tableau 3. Taux de reconnaissance en test sur les bases OSR et Pubfig.

- Une méthode appelée Qwise + LMNN + Fantope : cette méthode reprend le schéma Qwise + LMNN et y inclut le schéma de régularisation proposé dans (Law *et al.*, 2014). Cette régularisation utilise la notion de Fantope et généralise la régularisation par la norme nucléaire. Elle consiste à minimiser les plus petites valeurs singulières de la matrice SDP apprise afin d’obtenir un modèle de faible rang, et donc de limiter le surapprentissage.

Résultats quantitatifs : Le Tableau 3 présente les résultats obtenus avec les différentes méthodes. Nous utilisons pour toutes les méthodes un modèle gaussien par classe pour utiliser le même modèle de reconnaissance que (Parikh et Grauman, 2011), sauf pour "LMNN" et "Qwise+LMNN" pour lesquels un classifieur k -NN est utilisé car LMNN est optimisé pour être utilisé avec un classifieur k -NN. Notre algorithme donne de meilleures performances que les baselines, et l’information d’attributs relatifs et de classe ("Qwise + LMNN") gagne à être combinée sur la base PubFig. L’ajout du terme de régularisation par le Fantope améliore encore les résultats.

6. Conclusion

Dans cet article, nous avons proposé un modèle efficace d’apprentissage de distance de Mahalanobis qui exploite des contraintes entre quadruplets d’images. Nous montrons dans différents scénarii (tels que la détection de changements importants entre versions de page Web, et les attributs relatifs) qu’il est adapté pour incorporer de l’information riche de similarité.

Nous avons proposé une nouvelle méthode de détection de changement de page Web qui exploite des relations temporelles entre versions et détecte des régions sémantiques importantes. Dans le contexte des attributs relatifs, nous avons montré que les comparaisons entre paires d’images sont parfois limitées et peuvent être améliorées avec des comparaisons entre quadruplets d’images.

Remerciements Ce travail a été partiellement financé par le projet ANR VISIIR (ANR-13-CORD-0009).

7. Bibliographie

- Adar E., Teevan J., Dumais S., « Resonance on the web : web dynamics and revisitation patterns », *International Conference on Human Factors in Computing Systems (CHI)*, ACM, 2009.
- Ben Saad M., Gançarski S., « Archiving the Web using Page Changes Pattern : A Case Study », *ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, 2011.
- Cai D., Yu S., Wen J., Ma W., « VIPS : a Vision-based Page Segmentation Algorithm », *Microsoft Technical Report, MSR-TR-2003-79-2003*, 2003.
- Chapelle O., « Training a support vector machine in the primal », *Neural Computation*, 2007.
- Chechik G., Sharma V., Shalit U., Bengio S., « Large scale online learning of image similarity through ranking », *JMLR*, vol. 11, p. 1109-1135, 2010.
- Frome A., Singer Y., Malik J., « Image Retrieval and Classification Using Local Distance Functions », *NIPS*, 2006.
- Joachims T., « Optimizing search engines using clickthrough data », *SIGKDD*, 2002.
- Kumar N., Berg A., Belhumeur P., Nayar S., « Attribute and simile classifiers for face verification », *International Conference on Computer Vision (ICCV)*, 2009.
- Lampert C. H., Nickisch H., Harmeling S., « Learning to detect unseen object classes by between-class attribute transfer », *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.
- Law M. T., Thome N., Cord M., « Quadruplet-wise Image Similarity Learning », *International Conference on Computer Vision (ICCV)*, 2013.
- Law M. T., Thome N., Cord M., « Fantope Regularization in Metric Learning », *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.
- Law M. T., Thome N., Gançarski S., Cord M., « Structural and Visual Comparisons for Web Page Archiving », *ACM Symposium on Document Engineering (DocEng)*, 2012.
- Mignon A., Jurie F., « PCCA : A New Approach for Distance Learning from Sparse Pairwise Constraints », 2012.
- Oliva A., Torralba A., « Modeling the shape of the scene : A holistic representation of the spatial envelope », *International journal of computer vision (IJCV)*, vol. 42, n° 3, p. 145-175, 2001.
- Parikh D., Grauman K., « Relative Attributes », *International Conference on Computer Vision (ICCV)*, 2011.
- Pehlivan Z., Ben Saad M., Gançarski S., « Vi-DIFF : Understanding Web Pages Changes », *International Conference on Database and Expert Systems Applications (DEXA)*, 2010.
- Perrot M., Habrard A., Muselet D., Sebban M., « Modeling Perceptual Color Differences by Local Metric Learning », *Computer Vision–ECCV 2014*, Springer, p. 96-111, 2014.
- Song R., Liu H., Wen J., Ma W., « Learning block importance models for web pages », *WWW*, 2004.
- Weinberger K., Saul L., « Distance metric learning for large margin nearest neighbor classification », *JMLR*, vol. 10, p. 207-244, 2009.
- Xing E., Ng A., Jordan M., Russell S., « Distance metric learning, with application to clustering with side-information », *NIPS*, 2002.