Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence

Unified Constraint Propagation on Multi-View Data

Zhiwu Lu and Yuxin Peng*
Institute of Computer Science and Technology, Peking University, Beijing 100871, China
{luzhiwu,pengyuxin}@pku.edu.cn

Abstract

This paper presents a unified framework for intra-view and inter-view constraint propagation on multi-view data. Pairwise constraint propagation has been studied extensively, where each pairwise constraint is defined over a pair of data points from a single view. In contrast, very little attention has been paid to inter-view constraint propagation, which is more challenging since each pairwise constraint is now defined over a pair of data points from different views. Although both intra-view and inter-view constraint propagation are crucial for multi-view tasks, most previous methods can not handle them simultaneously. To address this challenging issue, we propose to decompose these two types of constraint propagation into semi-supervised learning subproblems so that they can be uniformly solved based on the traditional label propagation techniques. To further integrate them into a unified framework, we utilize the results of intra-view constraint propagation to adjust the similarity matrix of each view and then perform inter-view constraint propagation with the adjusted similarity matrices. The experimental results in cross-view retrieval have shown the superior performance of our unified constraint propagation.

Introduction

As an alternative type of supervisory information easier to access than the class labels of data points, pairwise constraints are widely used for different machine learning tasks in the literature. To effectively exploit pairwise constraints for machine learning, much attention has been paid to pairwise constraint propagation (Lu and Carreira-Perpinan 2008; Li, Liu, and Tang 2008; Yu and Shi 2004). Different from the method proposed in (Kamvar, Klein, and Manning 2003) which only adjusts the similarities between constrained data points, these constraint propagation approaches can propagate pairwise constraints to other similarities between unconstrained data points and thus achieve better results in most cases. More importantly, given that each pairwise constraint is actually defined over a pair of data points from a single view, these approaches can all be regarded as intra-view constraint propagation when multi-view data is concerned. Since we have to learn the relationships (must-link or cannot-link) between data points, intra-view constraint propagation is more challenging than the traditional label propagation (Zhou et al. 2004; Zhu, Ghahramani, and Lafferty 2003; Wang and Zhang 2008) whose goal is only to predict the labels of unlabeled data points.

However, besides intra-view pairwise constraints, we may also have easy access to inter-view pairwise constraints in multi-view tasks such as cross-view retrieval (Rasiwasia et al. 2010), where each pairwise constraint is defined over a pair of data points from different views (see Figure 1). In this case, inter-view pairwise constraints still specify the must-link or cannot-link relationships between data points. Since the similarity of two data points from different views are commonly unknown in practice, inter-view constraint propagation is significantly more challenging than intra-view constraint propagation. In fact, very little attention has been paid to inter-view constraint propagation for multi-view tasks in the literature. Although pairwise constraint propagation has been successfully applied to multi-view clustering in (Eaton, desJardins, and Jacob 2010; Fu et al. 2011), only intra-view pairwise constraints are propagated across different views. Here, it should be noted that these two constraint propagation methods *have actually ignored* the concept of inter-view pairwise constraints or the strategy of inter-view constraint propagation.

Since multi-view data can be readily decomposed into a series of two-view data, we focus on inter-view constraint propagation only across two views. However, such inter-view constraint propagation remains a rather challenging task. Fortunately, from a semi-supervised learning viewpoint, we can formulate both inter-view and intra-view constraint propagation as minimizing a regularized energy functional. Specifically, we first decompose the intra-view or inter-view constraint propagation problem into a set of independent semi-supervised learning (Zhou et al. 2004; Zhu, Ghahramani, and Lafferty 2003; Wang and Zhang 2008) subproblems. Through formulating these subproblems uniformly as minimizing a regularized energy functional, we thus develop efficient intra-view and inter-view constraint propagation algorithms based on the label propagation technique (Zhou et al. 2004). In summary, we succeed in giving a unified explanation of intra-view and inter-view constraint propagation from a semi-supervised learning viewpoint.

*Corresponding author.

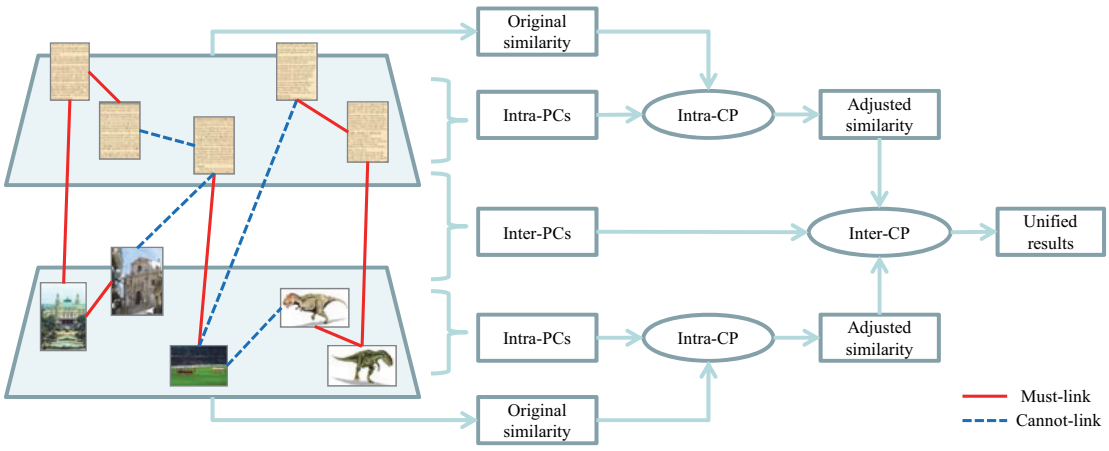

Figure 1: Illustration of the proposed framework for unified constraint propagation across two different views: text and image. Here, Intra-PCs and Inter-PCs denote respectively intra-view and inter-view pairwise constraints, while Intra-CP and Inter-CP denote respectively intra-view and inter-view constraint propagation.

Furthermore, to integrate intra-view and inter-view constraint propagation into a unified framework, we utilize the results of intra-view constraint propagation to adjust the similarity matrix of each view and then perform inter-view constraint propagation with the adjusted similarity matrices. The proposed framework for unified constraint propagation is illustrated in Figure 1. When multiple views refer to text, image, audio and so on, the output of our unified constraint propagation actually denotes the correlation between different media views. That is, our approach can be directly used for cross-view retrieval which has drawn much attention recently (Rasiwasia et al. 2010). For cross-view retrieval, it is not feasible to combine multiple views just as previous multi-view retrieval methods (Guillaumin, Verbeek, and Schmid 2010; Bruno, Moenne-Loccoz, and Marchand-Maillet 2008; Bekkerman and Jeon 2007; Snoek and Worring 2005). More notably, the two related methods (Eaton, desJardins, and Jacob 2010; Fu et al. 2011) for multi-view clustering are incompetent for cross-view retrieval.

Finally, the main contributions of the present work are summarized as follows:

- The intra-view and inter-view constraint propagation on multi-view data have been *uniformly explained* as minimizing a regularized energy functional from a semi-supervised learning viewpoint.
- The intra-view and inter-view constraint propagation have been *successfully integrated* into a unified framework for dealing with multi-view tasks.
- Although only tested in cross-view retrieval, our unified constraint propagation can be readily extended to other challenging multi-view tasks.

The remainder of this paper is organized as follows. In Section 2, we present the unified constraint propagation framework. In Section 3, our unified constraint propagation is applied to cross-view retrieval. In Section 4, we provide the experimental results to evaluate our unified constraint propagation. Finally, Section 5 gives the conclusions.

Unified Constraint Propagation

This section presents our unified constraint propagation on multi-view data. We first formulate intra-view constraint propagation as minimizing a regularized energy functional from a semi-supervised learning viewpoint. Furthermore, we similarly formulate the more challenging problem of inter-view constraint propagation as minimizing a regularized energy functional. Finally, we integrate these two types of constraint propagation into a unified framework.

Intra-View Constraint Propagation

Given a dataset $\mathcal{X} = \{x_1, ..., x_N\}$ from a single view, we denote a set of initial must-link constraints as $\mathcal{M} = \{(x_i, x_j) : l_i = l_j\}$ and a set of initial cannot-link constraints as $\mathcal{C} = \{(x_i, x_j) : l_i \neq l_j\}$, where l_i is the label of data point x_i. The goal of intra-view constraint propagation is to spread the effect of the two set of initial intra-view pairwise constraints throughout the entire dataset. In this paper, we denote the exhaustive set of propagated intra-view pairwise constraints as $F \in \mathcal{F}$, where $\mathcal{F} = \{F = \{f_{ij}\}_{N\times N} : 1 \leq i, j \leq N\}$. It should be noted that $f_{ij} > 0$ (or < 0) means (x_i, x_j) is a must-link (or cannot-link) constraint, with $|f_{ij}|$ denoting the confidence score of (x_i, x_j) being a must-link (or cannot-link) constraint. Hence, intra-view constraint propagation is equivalent to finding the best solution $F^* \in \mathcal{F}$ based on $\mathcal{M}$ and $\mathcal{C}$. We will elaborate it as follows.

Although it is difficult to directly find the best solution $F^* \in \mathcal{F}$ to intra-view constraint propagation, we can tackle this challenging problem by decomposing it into semi-supervised learning subproblems. More concretely, we first denote both $\mathcal{M}$ and $\mathcal{C}$ defined over the dataset $\mathcal{X}$ with a single matrix $Z = \{z_{ij}\}_{N\times N}$:

$$z_{ij} = \begin{cases} +1, & (x_i, y_j) \in \mathcal{M}; \\ -1, & (x_i, y_j) \in \mathcal{C}; \\ 0, & \text{otherwise.} \end{cases} \quad (1)$$

By making vertical and horizontal observations on such initial matrix Z (which stores the initial intra-view pair-

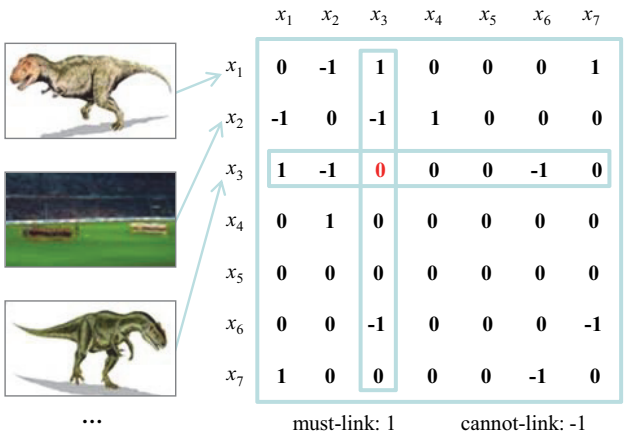

Figure 2: Illustration of Z. When we focus on a single data point (e.g. the third image x_3 here), the intra-view constraint propagation can be viewed as a two-class semi-supervised learning problem in both vertical and horizontal directions.

wise constraints), we further decompose the intra-view constraint propagation problem into semi-supervised learning subproblems, as illustrated in Figure 2.

Given a graph $\mathcal{G} = \{\mathcal{X}, W\}$ constructed over the dataset $\mathcal{X}$ with W being the edge weight matrix defined over the vertex set $\mathcal{X}$, we make use of graph-based method to solve these semi-supervised learning subproblems as follows:

$$\min_{F_v, F_h \in \mathcal{F}} \|F_v - Z\|_{fro}^2 + \mu \mathrm{tr}(F_v^T \mathcal{L} F_v) + \|F_h - Z\|_{fro}^2 + \mu \mathrm{tr}(F_h \mathcal{L} F_h^T) + \gamma \|F_v - F_h\|_{fro}^2, \quad (2)$$

where $\mu > 0$ (or $\gamma > 0$) denotes the regularization parameter, $\mathcal{L}$ denotes the normalized Laplacian matrix defined over the graph $\mathcal{G}$, $||\cdot||_{fro}$ denotes the Frobenius norm of a matrix, and $\mathrm{tr}(\cdot)$ denotes the trace of a matrix. The first and second terms of the above objective function are related to the *vertical* constraint propagation, while the third and fourth terms are related to the *horizontal* constraint propagation. The fifth term then ensures that the solutions of these types of constraint propagation are as approximate as possible. Let F_v^* and F_h^* be the best solutions of vertical and horizontal constraint propagation, respectively. The best solution of intra-view constraint propagation is defined as:

$$F^* = (F_v^* + F_h^*)/2. \quad (3)$$

As for the second and fourth terms, they are known as the energy functional (Zhu, Ghahramani, and Lafferty 2003) (or the smoothness measure) defined over $\mathcal{X}$ which has been widely used for graph-based semi-supervised learning. In summary, we have formulated intra-view constraint propagation as minimizing a regularized energy functional.

Let $\mathcal{Q}(F_v, F_h)$ denote the objective function in equation (2). The alternate optimization technique can adopted to solve $\min_{F_v, F_h} \mathcal{Q}(F_v, F_h)$ as follows: 1) Fix $F_h = F_h^*$, and perform vertical propagation by $F_v^* = \arg\min_{F_v} \mathcal{Q}(F_v, F_h^*)$; 2) Fix $F_v = F_v^*$, and perform horizontal propagation by $F_h^* = \arg\min_{F_h} \mathcal{Q}(F_v^*, F_h)$.

Vertical Propagation: When F_h is fixed at F_h^*, the solution of $\min_{F_v} \mathcal{Q}(F_v, F_h^*)$ can be found by solving

$$\frac{\partial \mathcal{Q}(F_v, F_h^*)}{\partial F_v} = 2(F_v - Z) + 2\mu \mathcal{L} F_v + 2\gamma(F_v - F_h^*) = 0,$$

which can be further transformed into

$$(I + \hat{\mu}\mathcal{L}) F_v = (1 - \beta) Z + \beta F_h^*, \quad (4)$$

where $\hat{\mu} = \mu/(1 + \gamma)$ and $\beta = \gamma/(1 + \gamma)$. Since $I + \hat{\mu}\mathcal{L}$ is positive definite, the above linear equation has a solution:

$$F_v^* = (I + \hat{\mu}\mathcal{L})^{-1}((1 - \beta) Z + \beta F_h^*). \quad (5)$$

However, this analytical solution is not efficient for large datasets, since matrix inverse has a time cost of $O(N^3)$. Fortunately, this analytical solution can also be *efficiently found using the label propagation technique* (Zhou et al. 2004) based on k-nearest neighbor (k-NN) graph.

Horizontal Propagation: When F_v is fixed at F_v^*, the solution of $\min_{F_h} \mathcal{Q}(F_v^*, F_h)$ can be found by solving

$$\frac{\partial \mathcal{Q}(F_v^*, F_h)}{\partial F_h} = 2(F_h - Z) + 2\mu F_h \mathcal{L} + 2\gamma(F_h - F_v^*) = 0,$$

which can be further transformed into

$$F_h (I + \hat{\mu}\mathcal{L}) = (1 - \beta) Z + \beta F_v^*, \quad (6)$$

which can also be efficiently solved using the label propagation technique (Zhou et al. 2004) based on k-NN graph.

Let W denote the weight matrix of the k-NN graph constructed over $\mathcal{X}$. The complete algorithm for intra-view constraint propagation is outlined as follows:

(1) Compute $S = D^{-\frac{1}{2}} W D^{-\frac{1}{2}}$, where D is a diagonal matrix with its entry (i, i) being the sum of row i of W;

(2) Initialize $F_v(0) = 0$, $F_h^* = 0$, and $F_h(0) = 0$;

(3) Iterate $F_v(t+1) = \alpha S F_v(t) + (1-\alpha)((1-\beta) Z + \beta F_h^*)$ until convergence at F_v^*, where $\alpha = \hat{\mu}/(1 + \hat{\mu})$;

(4) Iterate $F_h(t+1) = \alpha F_h(t) S + (1-\alpha)((1-\beta) Z + \beta F_v^*)$ until convergence at F_h^*;

(5) Iterate Steps (3)–(4) until the stopping condition is satisfied, and output the solution $F^* = (F_v^* + F_h^*)/2$.

Similar to (Zhou et al. 2004), the iteration in Step (3) converges to $F_v^* = (1 - \alpha)(I - \alpha S)^{-1}((1 - \beta) Z + \beta F_h^*)$, which is equal to the solution (5) given that $\alpha = \hat{\mu}/(1 + \hat{\mu})$ and $S = I - \mathcal{L}$. Moreover, in our later experiments, we find that the iterations in Steps (3)–(5) generally converge in very limited steps ($<$10). Finally, based on k-NN graph, our algorithm has a time cost of $O(kN^2)$ proportional to the number of all possible pairwise constraints. Hence, it can be considered to provide an efficient solution (note that even a simple assignment operator on F^* incurs a time cost of $O(N^2)$).

Inter-View Constraint Propagation

We have just provided a sound solution to the challenging problem of intra-view constraint propagation. However, this solution is limited to a single view. In this paper, we further consider even more challenging constraint propagation on multi-view data where each pairwise constraint is defined

over a pair of data points from different views. Since this problem can be readily decomposed into a series of two-view subproblems, we focus on inter-view constraint propagation on two-view data and then formulate it from a semi-supervised learning viewpoint.

Let $\{\mathcal{X}, \mathcal{Y}\}$ be a two-view dataset, where $\mathcal{X} = \{x_1, ..., x_N\}$ and $\mathcal{Y} = \{y_1, ..., y_M\}$. It should be noted that $\mathcal{X}$ may have a different data size from $\mathcal{Y}$ (i.e. $N \neq M$). As an example, a two-view dataset is shown in Figure 1, with image and text being the two views. For the two-view dataset $\{\mathcal{X}, \mathcal{Y}\}$, we can define a set of initial must-link constraints as $\mathcal{M} = \{(x_i, y_j) : l(x_i) = l(y_j)\}$ and a set of initial cannot-link constraints as $\mathcal{C} = \{(x_i, y_j) : l(x_i) \neq l(y_j)\}$, where $l(x_i)$ (or $l(y_j)$) is the label of $x_i \in \mathcal{X}$ (or $y_j \in \mathcal{Y}$). Here, x_i and y_j are assumed to share the same label set. If the class labels are not provided, the inter-view pairwise constraints can be defined only based on the correspondence between two views, which can be readily obtained from Web-based content such as Wikipedia articles.

Similar to our representation of the initial intra-view pairwise constraints, we can denote both $\mathcal{M}$ and $\mathcal{C}$ on the two-view dataset $\{\mathcal{X}, \mathcal{Y}\}$ with a single matrix $Z = \{z_{ij}\}_{N \times M}$:

$$z_{ij} = \begin{cases} +1, & (x_i, y_j) \in \mathcal{M}; \\ -1, & (x_i, y_j) \in \mathcal{C}; \\ 0, & \text{otherwise.} \end{cases} \tag{7}$$

Furthermore, by making vertical and horizontal observations on Z, we decompose the inter-view constraint propagation problem into semi-supervised learning subproblems, just as our interpretation of intra-view constraint propagation from a semi-supervised learning viewpoint. These subproblems can be similarly merged to a single optimization problem:

$$\min_{F_{\mathcal{X}}, F_{\mathcal{Y}}} \|F_{\mathcal{X}} - Z\|_{fro}^2 + \mu_{\mathcal{X}} \mathrm{tr}(F_{\mathcal{X}}^T \mathcal{L}_{\mathcal{X}} F_{\mathcal{X}}) + \|F_{\mathcal{Y}} - Z\|_{fro}^2 + \mu_{\mathcal{Y}} \mathrm{tr}(F_{\mathcal{Y}} \mathcal{L}_{\mathcal{Y}} F_{\mathcal{Y}}^T) + \gamma \|F_{\mathcal{X}} - F_{\mathcal{Y}}\|_{fro}^2, \tag{8}$$

where $\mu_{\mathcal{X}}$ (or $\mu_{\mathcal{Y}}$) denotes the regularization parameter for $\mathcal{X}$ (or $\mathcal{Y}$), and $\mathcal{L}_{\mathcal{X}}$ (or $\mathcal{L}_{\mathcal{Y}}$) denotes the normalized Laplacian matrix defined over $\mathcal{X}$ (or $\mathcal{Y}$). The second and fourth terms of the above equation denote the energy functional (Zhu, Ghahramani, and Lafferty 2003) (or the smoothness measure) defined over $\mathcal{X}$ and $\mathcal{Y}$, respectively. In summary, we have also formulated inter-view constraint propagation as minimizing a regularized energy functional.

It should be noted that the alternate optimization technique can be adopted to solve the above inter-view constraint propagation problem just as what we have done for equation (2). Let $W_{\mathcal{X}}$ (or $W_{\mathcal{Y}}$) denote the weight matrix of the k-NN graph constructed on $\mathcal{X}$ (or $\mathcal{Y}$). The complete algorithm for inter-view constraint propagation is summarized as follows:

(1) Compute two matrices $S_{\mathcal{X}} = D_{\mathcal{X}}^{-1/2} W_{\mathcal{X}} D_{\mathcal{X}}^{-1/2}$ and $S_{\mathcal{Y}} = D_{\mathcal{Y}}^{-1/2} W_{\mathcal{Y}} D_{\mathcal{Y}}^{-1/2}$, where $D_{\mathcal{X}}$ (or $D_{\mathcal{Y}}$) is a diagonal matrix with its i-th diagonal entry being the sum of the i-th row of $W_{\mathcal{X}}$ (or $W_{\mathcal{Y}}$);

(2) Initialize $F_{\mathcal{X}}(0) = 0$, $F_{\mathcal{Y}}^* = 0$, and $F_{\mathcal{Y}}(0) = 0$;

(3) Iterate $F_{\mathcal{X}}(t+1) = \alpha_{\mathcal{X}} S_{\mathcal{X}} F_{\mathcal{X}}(t) + (1-\alpha_{\mathcal{X}})((1-\beta)Z + \beta F_{\mathcal{Y}}^*)$ until convergence at $F_{\mathcal{X}}^*$, where $\alpha_{\mathcal{X}} = \hat{\mu}_{\mathcal{X}}/(1 + \hat{\mu}_{\mathcal{X}})$ with $\hat{\mu}_{\mathcal{X}} = \mu_{\mathcal{X}}/(1+\gamma)$, and $\beta = \gamma/(1+\gamma)$;

(4) Iterate $F_{\mathcal{Y}}(t+1) = \alpha_{\mathcal{Y}} F_{\mathcal{Y}}(t) S_{\mathcal{Y}} + (1-\alpha_{\mathcal{Y}})((1-\beta)Z + \beta F_{\mathcal{X}}^*)$ until convergence at $F_{\mathcal{Y}}^*$, where $\alpha_{\mathcal{Y}} = \hat{\mu}_{\mathcal{Y}}/(1+\hat{\mu}_{\mathcal{Y}})$ with $\hat{\mu}_{\mathcal{Y}} = \mu_{\mathcal{Y}}/(1+\gamma)$;

(5) Iterate Steps (3)–(4) until the stopping condition is satisfied, and output the solution $F^* = (F_{\mathcal{X}}^* + F_{\mathcal{Y}}^*)/2$.

In our later experiments, we find that the iterations in Steps (3)–(5) generally converge in very limited steps (<10). Moreover, based on k-NN graphs, the above algorithm has a time cost of $O(kNM)$ proportional to the number of all possible pairwise constraints. Hence, we consider that it can provide an efficient solution (note that even a simple assignment operator on F^* incurs a time cost of $O(NM)$).

The Unified Algorithm

Although we have given a unified explanation of intra-view and inter-view constraint propagation from a semi-supervised learning viewpoint, these two types of constraint propagation are completely separate when applied to multi-view tasks. To exploit the initial pairwise constraints most effectively, we propose to integrate them into a unified framework. That is, we first adjust the similarity matrix of each view using the results of intra-view constraint propagation and then perform inter-view constraint propagation with the adjusted similarity matrices.

Let $F^* = \{f_{ij}^*\}_{N \times N}$ be the output of our intra-view constraint propagation for a view. The original normalized weight matrix W (i.e. $0 \leq w_{ij} \leq 1$) of the graph constructed for this view can be adjusted as (Lu and Ip 2010):

$$\tilde{w}_{ij} = \begin{cases} 1 - (1 - f_{ij}^*)(1 - w_{ij}), & f_{ij}^* \geq 0; \\ (1 + f_{ij}^*) w_{ij}, & f_{ij}^* < 0. \end{cases} \tag{9}$$

We then use $\tilde{W} = \{\tilde{w}_{ij}\}_{N \times N}$ as the new weight (or similarity) matrix. By such similarity adjustment, we can derive two new similarity matrices $\tilde{W}_{\mathcal{X}}$ and $\tilde{W}_{\mathcal{Y}}$ for a two-view dataset $\{\mathcal{X}, \mathcal{Y}\}$ from the original $W_{\mathcal{X}}$ and $W_{\mathcal{Y}}$.

Finally, we perform inter-view constraint propagation using the two new similarity matrices. The unified algorithm is summarized as follows:

(1) Perform intra-view constraint propagation for the two views (i.e. $\mathcal{X}$ and $\mathcal{Y}$), respectively;

(2) Derive the new similarity matrices $\tilde{W}_{\mathcal{X}}$ and $\tilde{W}_{\mathcal{Y}}$ using the results of intra-view constraint propagation;

(3) Perform inter-view constraint propagation with the new similarity matrices $\tilde{W}_{\mathcal{X}}$ and $\tilde{W}_{\mathcal{Y}}$;

(4) Output the unified propagated results.

The flowchart of the above unified constraint propagation (UCP) algorithm has been illustrated in Figure 1.

Application to Cross-View Retrieval

When multiple views refer to text, image, audio and so on, the output of our unified constraint propagation actually can be viewed as the correlation between different media views. As we have mentioned, given the output $F^* = \{f_{ij}^*\}_{N \times M}$ of our unified constraint propagation, (x_i, y_j) denotes a must-link (or cannot-link) constraint if $f_{ij}^* > 0$ (or < 0). Considering the inherent meanings of must-link and cannot-link

constraints, we can state that: x_i and y_j are "positively correlated" if $f^*_{ij} > 0$, while they are "negatively correlated" if $f^*_{ij} < 0$. Hence, we can view f^*_{ij} as the correlation coefficient between x_i and y_j. The distinct advantage of such interpretation of F^* as a correlation measure is that F^* can thus be used for ranking on $\mathcal{Y}$ given a query x_i or ranking on $\mathcal{X}$ given a query y_j. In fact, this is just the goal of cross-view retrieval which has drawn much attention recently (Rasiwasia et al. 2010). That is, such task can be directly handled by our unified constraint propagation.

In this paper, we focus on a special case of cross-view retrieval, i.e. only text and image views are considered. In this case, cross-view retrieval is somewhat similar to automatic image annotation (Li and Wang 2003; Feng, Manmatha, and Lavrenko 2004; Fan et al. 2011) and image caption generation (Farhadi et al. 2010; Kulkarni et al. 2011; Ordonez, Kulkarni, and Berg 2012), since these three tasks all aim to learn the relations between the text and image views. However, even if only text and image views are considered, cross-view retrieval is still quite different from automatic image annotation and image caption generation. More concretely, automatic image annotation relies on very limited types of textual representations and mainly associates images only with textual keywords, while cross-view retrieval is designed to deal with much more richly annotated data, motivated by the ongoing explosion of Web-based multimedia content such as news archives and Wikipedia pages. Similar to cross-view retrieval, image caption generation can also deal with more richly annotated data (i.e. captions) with respect to the textual keywords concerned in automatic image annotation. However, this challenging task tends to model image captions as sentences by exploiting certain prior knowledge (e.g. the <object, action, scene> triplets used in (Farhadi et al. 2010; Kulkarni et al. 2011)), different from cross-view retrieval that focuses on associating images with complete text articles using no prior knowledge from the text view (any general textual representations are applicable actually once their similarities are provided).

In the context of cross-view retrieval, one notable recent work is (Rasiwasia et al. 2010) which has investigated two hypotheses for problem formulation: 1) there is a benefit to explicitly modeling the correlation between text and image views, and 2) this modeling is more effective with higher levels of abstraction. More concretely, the correlation between the two views is learned with canonical correlation analysis (CCA) (Hotelling 1936) and abstraction is achieved by representing text and image at a more general semantic level. However, two separate steps, i.e. correlation analysis and semantic abstraction, are involved in this modeling, and the use of semantic abstraction after CCA seems rather ad hoc. Fortunately, this problem can be completely addressed by our unified constraint propagation. The semantic information (e.g. class labels) associated with images and text can be used to define the initial must-link and cannot-link constraints based on the training dataset, while the correlation between text and image views can be explicitly learnt by the proposed algorithm in Section 2. That is, the correlation analysis and semantic abstraction has been successfully integrated in our unified constraint propagation framework. More importantly, our later experimental results have demonstrated the effectiveness of such integration as compared to the recent work (Rasiwasia et al. 2010).

Experimental Results

In this section, our unified constraint propagation (UCP) algorithm is evaluated in the challenging application of cross-view retrieval. We focus on comparing our UCP algorithm with the state-of-the-art approach (Rasiwasia et al. 2010), since they both consider not only correlation analysis (CA) but also semantic abstraction (SA) for text and image views. Moreover, we also make comparison with another two closely related approaches that integrate CA and SA for cross-view retrieval similar to (Rasiwasia et al. 2010) but perform correlation analysis by partial least squares (PLS) (Wold 1985) and cross-modal factor analysis (CFA) (Li et al. 2003) instead of CCA, respectively. In the following, these two CA+SA approaches are denoted as CA+SA (PLS) and CA+SA (CFA), while the state-of-the-art approach (Rasiwasia et al. 2010) is denoted as CA+SA (CCA).

Experimental Setup

We select two different datasets for performance evaluation. The first one is a Wikipedia benchmark dataset (Rasiwasia et al. 2010), which contains a total of 2,866 documents derived from Wikipedia's "featured articles". Each document is actually a text-image pair, annotated with a label from the vocabulary of 10 semantic classes. This benchmark dataset (Rasiwasia et al. 2010) is split into a training set of 2,173 documents and a test set of 693 documents. Moreover, the second dataset consists of totally 8,564 documents crawled from the photo sharing website Flickr. The image and text views of each document denote a photo and a set of tags provided by the users, respectively. Although such text presentation does not take a free form as that for the Wikipedia dataset, it is rather noisy since many of the tags may be incorrectly annotated by the users. This Flickr dataset is organized into 11 semantic classes. We split it into a training set of 4,282 documents and a test set of the same size.

For the above two datasets, we take the same strategy as (Rasiwasia et al. 2010) to generate both text and image representation. More concretely, in the Wikipedia dataset, the text representation for each document is derived from a latent Dirichlet allocation model (Blei, Ng, and Jordan 2003) with 10 latent topics, while the image representation is based on a bag-of-words model with 128 visual words learnt from the extracted SIFT descriptors (Lowe 2004), just as (Rasiwasia et al. 2010). Moreover, for the Flickr dataset, we generate the text and image representation similarly, and the main difference is that we select a relatively large visual vocabulary (of the size 2,000) and refine the noisy textual vocabulary to the size 1,000 as a preprocessing step. For both text and image representation, the normalized correlation is used as the similarity measure just as (Rasiwasia et al. 2010).

In our experiments, the initial intra-view and inter-view pairwise constraints for our UCP algorithm are derived from the class labels of the training documents of each dataset. The performance of our UCP algorithm is evaluated on the

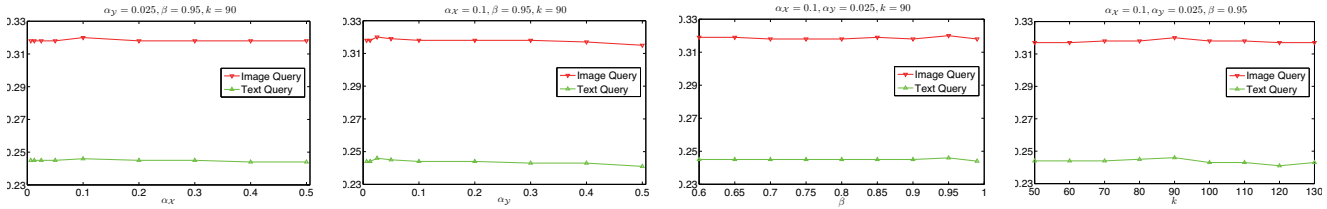

Figure 3: The retrieval results by cross-validation on the training set of the Wikipedia dataset for our Inter-CP algorithm.

Table 1: The retrieval results on the test set of the Wikipedia dataset measured by the MAP scores.

Methods	Image Query	Text Query	Average
CA+SA (PLS)	0.250	0.190	0.220
CA+SA (CFA)	0.272	0.221	0.247
CA+SA (CCA)	0.277	0.226	0.252
Inter-CP (Ours)	0.329	0.256	0.293
UCP (Ours)	**0.337**	**0.260**	**0.299**

Table 2: The retrieval results on the test set of the Flickr dataset measured by the MAP scores.

Methods	Image Query	Text Query	Average
CA+SA (PLS)	0.201	0.168	0.185
CA+SA (CFA)	0.252	0.231	0.242
CA+SA (CCA)	0.280	0.263	0.272
Inter-CP (Ours)	0.495	0.483	0.489
UCP (Ours)	**0.521**	**0.499**	**0.510**

test set. Here, two tasks of cross-view retrieval are considered: text retrieval using an image query, and image retrieval using a text query. In the following, these two tasks are denoted as "Image Query" and "Text Query", respectively. For each task, the retrieval results are measured with mean average precision (MAP) which has been widely used in the image retrieval literature (Rasiwasia, Moreno, and Vasconcelos 2007; Guillaumin, Verbeek, and Schmid 2010).

Let $\mathcal{X}$ denote the text representation and $\mathcal{Y}$ denote the image representation for each dataset. For our UCP algorithm (i.e. Intra-CP+Inter-CP), we construct k-NN graphs on $\mathcal{X}$ and $\mathcal{Y}$ with the same k. The parameters of our UCP algorithm can be selected by fivefold cross-validation on the training set. For example, according to Figure 3, we set the parameters of Inter-CP on the Wikipedia dataset as: $\alpha_{\mathcal{X}} = 0.1$, $\alpha_{\mathcal{Y}} = 0.025$, $\beta = 0.95$, and $k = 90$. It is noteworthy that our Inter-CP algorithm is not sensitive to these parameters. Moreover, the parameters of Intra-CP can be similarly set to their respective optimal values. To summarize, we have selected the best values for all the parameters of our UCP algorithm by cross-validation on the training set. For fair comparison, we take the same parameter selection strategy for other closely related algorithms.

Retrieval Results

The cross-view retrieval results on the two datasets are listed in Tables 1 and 2, respectively. The immediate observation is that we can achieve the best results when both intra-view and inter-view constraint propagation are considered (i.e. UCP here). This means that our UCP can most effectively exploit the initial supervisory information provided for cross-view retrieval. As compared to the three CA+SA approaches by semantic abstraction after correlation analysis (via PLS, CFA, or CCA), our Inter-CP can seamlessly integrate these two separate steps and then lead to much better results. Moreover, the effectiveness of Intra-CP is verified by the comparison between Inter-CP and UCP, especially on the Flickr dataset. In summary, we have verified the effectiveness of both Inter-CP and Intra-CP for cross-view retrieval.

It should be noted that our Inter-CP (or Intra-CP) algorithm can be considered to provide an efficient solution, since it has a time cost proportional to the number of all possible pairwise constraints. This is also verified by our observations in the experiments. For example, the running time taken by CA+SA (CCA, CFA or PLS), Inter-CP, and UCP on the Wikipedia dataset is 10, 24, and 55 seconds, respectively. Here, we run all the algorithms (Matlab code) on a computer with 3GHz CPU and 32GB RAM. Since our UCP (or Inter-CP) algorithm leads to significant better results in cross-view retrieval, we prefer it to CA+SA in practice, regardless of its relatively larger time cost.

Conclusions

We have investigated the challenging problems of intra-view and inter-view constraint propagation on multi-view data. By decomposing these two problems into independent semi-supervised learning subproblems, we have uniformly formulated them as minimizing a regularized energy functional. More importantly, the semi-supervised learning subproblems can be solved efficiently using label propagation with k-NN graphs. To further integrate them into a unified framework, we utilize the results of intra-view constraint propagation to adjust the similarity matrix of each view for inter-view constraint propagation. The experimental results in cross-view retrieval have shown the superior performance of our unified constraint propagation. For future work, our method will be extended to other multi-view tasks.

Acknowledgements

This work was supported by National Natural Science Foundation of China under Grants 61073084 and 61202231, Beijing Natural Science Foundation of China under Grants 4122035 and 4132037, Ph.D. Programs Foundation of Ministry of Education of China under Grants 20120001110097 and 20120001120130, and National Hi-Tech Research and Development Program (863 Program) of China under Grant 2012AA012503.

References

Bekkerman, R., and Jeon, J. 2007. Multi-modal clustering for multimedia collections. In *Proc. CVPR*, 1–8.

Blei, D.; Ng, A.; and Jordan, M. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research* 3:993–1022.

Bruno, E.; Moenne-Loccoz, N.; and Marchand-Maillet, S. 2008. Design of multimodal dissimilarity spaces for retrieval of video documents. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 30(9):1520–1533.

Eaton, E.; desJardins, M.; and Jacob, S. 2010. Multi-view clustering with constraint propagation for learning with an incomplete mapping between views. In *Proceedings of ACM International Conference on Information and Knowledge Management*, 389–398.

Fan, J.; Shen, Y.; Yang, C.; and Zhou, N. 2011. Structured max-margin learning for inter-related classifier training and multilabel image annotation. *IEEE Transactions on Image Processing* 20(3):837–854.

Farhadi, A.; Hejrati, M.; Sadeghi, M.; Young, P.; Rashtchian, C.; Hockenmaier, J.; and Forsyth, D. 2010. Every picture tells a story: Generating sentences from images. In *Proc. ECCV*, volume 4, 15–29.

Feng, S.; Manmatha, R.; and Lavrenko, V. 2004. Multiple Bernoulli relevance models for image and video annotation. In *Proc. CVPR*, volume 2, 1002–1009.

Fu, Z.; Ip, H.; Lu, H.; and Lu, Z. 2011. Multi-modal constraint propagation for heterogeneous image clustering. In *Proceedings of ACM Internatial Conference on Multimedia*, 143–152.

Guillaumin, M.; Verbeek, J.; and Schmid, C. 2010. Multimodal semi-supervised learning for image classification. In *Proc. CVPR*, 902–909.

Hotelling, H. 1936. Relations between two sets of variates. *Biometrika* 28(3-4):321–377.

Kamvar, S.; Klein, D.; and Manning, C. 2003. Spectral learning. In *Proceedings of International Joint Conferences on Artificial Intelligence*, 561–566.

Kulkarni, G.; Premraj, V.; Dhar, S.; Li, S.; Choi, Y.; Berg, A.; and Berg, T. 2011. Baby talk: Understanding and generating simple image descriptions. In *Proc. CVPR*, 1601–1608.

Li, J., and Wang, J. 2003. Automatic linguistic indexing of pictures by a statistical modeling approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 25(9):1075–1088.

Li, D.; Dimitrova, N.; Li, M.; and Sethi, I. K. 2003. Multimedia content processing through cross-modal association. In *Proceedings of ACM Internatial Conference on Multimedia*, 604–611.

Li, Z.; Liu, J.; and Tang, X. 2008. Pairwise constraint propagation by semidefinite programming for semi-supervised classification. In *Proceedings of International Conference on Machine Learning*, 576–583.

Lowe, D. 2004. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision* 60(2):91–110.

Lu, Z., and Carreira-Perpinan, M. 2008. Constrained spectral clustering through affinity propagation. In *Proc. CVPR*, 1–8.

Lu, Z., and Ip, H. 2010. Constrained spectral clustering via exhaustive and efficient constraint propagation. In *Proc. ECCV*, volume 6, 1–14.

Ordonez, V.; Kulkarni, G.; and Berg, T. 2012. Im2Text: Describing images using 1 million captioned photographs. In *Advances in Neural Information Processing Systems 24*, 1143–1151.

Rasiwasia, N.; Costa Pereira, J.; Coviello, E.; Doyle, G.; Lanckriet, G.; Levy, R.; and Vasconcelos, N. 2010. A new approach to cross-modal multimedia retrieval. In *Proceedings of ACM Internatial Conference on Multimedia*, 251–260.

Rasiwasia, N.; Moreno, P.; and Vasconcelos, N. 2007. Bridging the gap: Query by semantic example. *IEEE Transactions on Multimedia* 9(5):923–938.

Snoek, C., and Worring, M. 2005. Multimodal video indexing: A review of the state-of-the-art. *Multimedia Tools and Applications* 25(1):5–35.

Wang, F., and Zhang, C. 2008. Label propagation through linear neighborhoods. *IEEE Transactions on Knowledge and Data Engineering* 20(1):55–67.

Wold, H. 1985. Partial least squares. In Kotz, S., and Johnson, N., eds., *Encyclopedia of Statistical Sciences*. New York: Wiley. 581–591.

Yu, S., and Shi, J. 2004. Segmentation given partial grouping constraints. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 26(2):173–183.

Zhou, D.; Bousquet, O.; Lal, T.; Weston, J.; and Schölkopf, B. 2004. Learning with local and global consistency. In *Advances in Neural Information Processing Systems 16*, 321–328.

Zhu, X.; Ghahramani, Z.; and Lafferty, J. 2003. Semi-supervised learning using Gaussian fields and harmonic functions. In *Proceedings of International Conference on Machine Learning*, 912–919.