Graph Based Multi-Modality Learning*

Hanghang Tong[1], Jingrui He[1], Mingjing Li[2], Changshui Zhang[1], Wei-Ying Ma[2]
[1]Automation Department, Tsinghua University, Beijing 100084, China
+86-10-62782447
{walkstar98, hejingrui98}@mails.tsinghua.edu.cn
zcs@mail.tsinghua.edu.cn
[2]Microsoft Research Asia, 49 Zhichun Road, Beijing 100080, China
+86-10-62617711
{mjli, wyma}@microsoft.com

ABSTRACT
To better understand the content of multimedia, a lot of research efforts have been made on how to learn from multi-modal feature. In this paper, it is studied from a graph point of view: each kind of feature from one modality is represented as one independent graph; and the learning task is formulated as inferring from the constraints in every graph as well as supervision information (if available). For semi-supervised learning, two different fusion schemes, namely linear form and sequential form, are proposed. For each scheme, it is derived from optimization point of view; and further justified from two sides: similarity propagation and Bayesian interpretation. By doing so, we reveal the regular optimization nature, transductive learning nature as well as prior fusion nature of the proposed schemes, respectively. Moreover, the proposed method can be easily extended to unsupervised learning, including clustering and embedding. Systematic experimental results validate the effectiveness of the proposed method.

Categories and Subject Descriptors
H.3.1 [**INFORMATION STORAGE AND RETRIEVAL**]: Content Analysis and Index.

General Terms: Algorithms, Theory, Experimentation.

Keywords: Multi-modality analysis; graph model; regularized optimization; similarity propagation; Bayesian interpretation.

1. INTRODUCTION
In multimedia content analysis, much research effort has been made to utilize the multi-modal feature to better understand the multimedia content in recent years, which benefits from the following fact: the representation for the data point can be naturally split into two or more independent modalities [1, 3, 5, 6, 8, 11, 24, 25]. For example, Web page can be represented by its plain text as well as the anchor text; video can be represented by visual, audio, and caption track; digital image can be represented by color feature and texture feature; Web image can be represented by content feature and its text annotation. A lot of research demonstrates that by properly fusing the evidence from each modality, better understanding could be achieved than only using one modality or simply treating all representation as one modality.

According to the different learning tasks, existing work in multi-modality learning can be classified into three categories: supervised learning, semi-supervised learning, and un-supervised learning.

Tradition methods mainly focus on supervised learning task. According to at which level fusion takes place, they can be further classified into two categories: fusion at feature level and fusion at output level [8]. The work in [5] belongs to the first category, in which textual and visual features are concatenated into one single index vector for Web image retrieval. On the other hand, it has been recognized that fusion on output level generally outperforms the former [6, 8, 11]. Many fusion strategies can be adopted in this case, including linear combination, min-max aggregation, voting production combination etc [14, 19]. Among them, one most widely used strategy is linear combination [11, 14]. However, as pointed out by [25], linear combination has its own theoretical limitation. To address this issue, by treating the output of each classifier as a new kind of feature, the authors in [24] proposed a non-linear fusion method named super-kernel. However, the unlabelled data is still not explored.

To leverage the unlabelled data in the training stage, semi-supervised learning has been applied into multi-modality learning. One of the most widely used methods in this category is Co-Train [3]. The authors in [3] justified Co-Train in the Probably Approximately Correct (PAC) framework, provided that two modalities are compatible and un-correlated. The authors in [7] applied Co-Train in Web image annotation and retrieval. However, in real applications, the two assumptions of Co-Train (compatibility and un-correlation) are not always satisfied. An important variant of Co-Train is Co-EM [9, 17], which uses the hypothesis in one modality to probabilistically label the sample in the other modality. The authors in [17] argue that Co-EM is

MM'05, November 6–11, 2005, Singapore.

* This work was performed at Microsoft Research Asia.

closer to theoretical assumptions in [3] than Co-Train. However, as pointed out by [26], Co-EM is just a technical design and its convergence is not proved.

To cluster objects with heterogeneous features, some researchers have extended existing clustering methods to multi-modality version. For example, the authors in [23] proposed a reinforcement clustering algorithm; the authors in [13] extended existing DBSCAN algorithm to multi-modality; the authors in [2] extended EM to Multi-View EM based on mixture model. Compared with supervised learning, unsupervised learning in multi-modality seems not to be fully explored.

On a more general level, there are two key issues in multi-modality learning for any learning task: 1) how to learn within each modality; 2) how to fuse the evidence across each modality. It can be seen that most of existing work in multi-modality learning is based on vector model.

More recently, graph based learning method has attracted more and more research attention in both learning society and multimedia community. For unsupervised learning task, spectral clustering has shown its superiority in many applications. For example, the authors in [27] proposed Locality Preserving Cluster (LPC) algorithm for clustering images and demonstrated its advantage over traditional methods. In [4], the authors applied Laplacian Eigen-Map [1] to hierarchically cluster Web image search result. For semi-supervised learning task, the authors in [10] applied a recent developed manifold ranking algorithm in content based image retrieval (CBIR) in the scenario of query by example (QBE). By exploring the relationship among all images in the database, the authors showed that it outperforms the state-of-art techniques in CBIR. The authors in [20] further extend the manifold ranking algorithm to image retrieval in the scenario of query by keyword (QBK).

Inspired by the success of [4, 10, 20, 27], in this paper, we take a further step on graph based methods and explore their extensions in multi-modality learning task. From graph point of view, each kind of feature from one modality is represented as one independent graph and the learning task is formulated as inferring from the constraints in every graph as well as supervision information (if available).

For semi-supervised learning, both classification and retrieval in QBK are considered. Based on different optimization strategies, two different fusion schemes, namely linear form and sequential form, are proposed. While in the linear form, all the constraints are fused simultaneously, they are considered sequentially in the sequential form. For each scheme, both closed solution and iterative solution are developed, providing the later converges to the former. To reveal the transductive nature as well as prior probability fusion nature, the proposed schemes are further justified from two sides: similarity propagation and Bayesian interpretation, respectively. By doing so, we will show that the difference between linear form and sequential form actually comes from 1) the different optimization strategy they adopt; 2) the different manner they spread and fuse similarity through graphs; and 3) the different way they fuse the prior probability.

On the other hand, the proposed method can be viewed as similarity matrix learning or graph Laplacian learning from multi-modality. Thus, by feeding the learnt matrix into some existing spectral clustering or embedding algorithms, the proposed method can be naturally extended to un-supervised learning.

The main contribution of this paper is summarized as follows:

1) Make a systematic investigation on graph based methods in terms of their extension in multi-modality learning. Both semi-supervised and un-supervised learning are investigated;
2) For semi-supervised learning, propose two different schemes. For each scheme, it is derived from optimization point of view and further justified from two sides: similarity propagation and Bayesian interpretation;
3) For un-supervised learning, extend a spectral clustering algorithm to multi-modality; extend a spectral embedding algorithm to multi-modality.

The organization of this paper is as follows. In Section 2, we make a short review on the related work. The proposed method for semi-supervised learning task is presented in Section 3. Its extension to unsupervised task is provided in Section 4. In Section 5, we provide systematic experimental results which demonstrate the effectiveness of our method. Finally, we conclude the paper in Section 6.

2. RELATED WORK

2.1 Manifold Ranking Algorithm

The manifold ranking algorithm is a graph based semi-supervised learning algorithm [30, 31]. It has two versions for different tasks: to rank data points and to predict the labels of unlabeled points.

For the task of predicting the labels of unlabeled data points, it can be formulated as: given a set of points $\chi = \{x_1, \ldots, x_l, x_{l+1}, \ldots, x_n\} \subset \mathbb{R}^m$ and a label set $\zeta = \{1, \cdots, c\}$, the first l points x_i $(i \le l)$ are labeled as $y_i \in \zeta$; and the remaining points x_u $(l+1 \le u \le n)$ are to be labeled.

Define a $n \times c$ labeling matrix $Y = [Y_1, \cdots, Y_c]$ with $Y_{ij} = 1$ if x_i is labeled as $y_i = j$ and $Y_{ij} = 0$ otherwise; define a $n \times c$ matrix F corresponding to a classification on the dataset χ by labeling each point x_i with $y_i = \arg\max_{j \le c} F_{ij}$. The procedure of predicting labels can be summarized as the follows [30]:

Algorithm 1 Manifold ranking algorithm

1. Let $W = (W_{ij}, i, j = 1, \cdots, n)$ an $n \times n$ affinity matrix.
2. Symmetrically normalize W by $S = D^{-1/2} W D^{-1/2}$ in which D is the diagonal matrix with (i,i)-element equal to the sum of the ith row of W.
3. Iterate $F(t+1) = \alpha S F(t) + (1-\alpha) Y$ until convergence, where α is a parameter in $[0,1)$ and $F(0) = Y$.
4. Let F^* denote the limit of the sequence $\{F(t)\}$. Label each point x_i with $y_i = \arg\max_{j \le c} F^*_{ij}$.

An intuitive description of the above algorithm is: a weighted graph is first formed which takes each data point as a vertex; a positive score is assigned to each label while zero to the remaining points; all the points then spread their scores to the nearby points via the weighted graph; the spread process is repeated until a global stable state is reached, and all the points will have their own scores according to which they will be labeled.

2.2 Application in Image Retrieval

In [10, 20], the authors have applied the above algorithm to image retrieval in the scenario of QBE and QBK, respectively. In both works, the authors showed that their algorithms outperform the state-of-art techniques. The key points of [10, 20] are briefly summarized as follows:

- In the initial query stage in the scenario of QBE, there is only one query in the label set (in this case, F is an $n\times 1$ ranking vector). The resultant ranking score of an unlabeled image is in proportion to the probability that it is relevant to the query, with large ranking score indicating high probability. For QBK, the keyword model, indicating the relevant score of each image with each keyword, is constructed from the initial labels and step 4 in algorithms 1 is ignored for soft annotation purpose. The initial retrieval result is given by sorting the image in the descending order of its relevant score with respect to the query.
- In relevance feedback (both in QBE and QBK), if the user only marks relevant examples, the algorithm can be easily generalized by adding these newly labeled images into the query set; on the other hand, if examples of both labels are available, they are treated differently: relevant images are also added to the query set, while for irrelevant images, the authors designed three schemes based on the observation that positive examples should make more contribution to the final ranking score than negative ones.
- To maximally improve the ranking result, the authors also developed three active learning methods for selecting images in each round of relevance feedback. Namely, 1) to select the most positive images; 2) to select the most informative images; and 3) to select the most positive and inconsistent images.

3. GRAPH BASED SEMI-SUPERVISED LEARNING IN MULTI-MODALITY

First, we address the graph based semi-supervised learning in multi-modality, including classification and retrieval in the scenario of QBK. After a brief statement of notation and problem definition, we will propose two different fusion schemes from optimization point of view, and further justify them from both similarity propagation point of view and Bayesian interpretation.

3.1 Notation and Problem Definition

We use the same notations as those in Algorithm 1, except that each data point has more than one kind of feature (one modality for one kind of feature). Without losing generality, we suppose each data point x_i has two kinds of feature: $x_i = \langle x_i^a, x_i^b \rangle$ $(i = 1,2,\cdots,n)$, where x_i^a and x_i^b denote the feature vector constructed from modality a and b, respectively.

Let $W^a = \left(W_{ij}^a, i,j = 1,2,\cdots,n\right)$ be an $n\times n$ affinity matrix constructed from $x_i^a (i = 1,2,\cdots,n)$, where W_{ij}^a denotes the similarity between x_i and x_j measured from modality a. Normalize W^a by $S^a = (D^a)^{-1/2} W^a (D^a)^{-1/2}$, where D^a is the diagonal matrix with (i,i)-element equal to the sum of the ith row of W^a;

Let W^b, D^b, and S^b be defined similarly as above, except that they are constructed from modality b;

Let Y be an $n\times c$ labeling matrix with $Y_{ij} = 1$ if x_i is labeled as $y_i = j$ and $Y_{ij} = 0$ otherwise; Y_i is a $1\times c$ labeling vector for data i;

Let F be an $n\times c$ vectorial function, where F_{ij} denotes the relevance of data i belonging to class j, with a larger value indicating higher relevance. F_i is a $1\times c$ classification vector for data i.

With the above notation, the learning task is to infer the vectorial function F from W^a, W^b and Y as Eq.1.

$$\{(W^a, D^a, S^a); (W^b, D^b, S^b); Y\} \to F \tag{1}$$

Once F is obtained, a classification decision on the dataset χ can be made by labeling each point x_i with $y_i = \arg\max_{j\le c} F_{ij}$. For retrieval task in the scenario of QBK, the relevance score of the data point x_i with respect to the query q is given by $S_i = F_{iq}$. The initial retrieval result is given by sorting the data points in the decreasing order of their relevance scores.

To best fuse the information of S^a, S^b and Y to improve classification or retrieval performance, a 'good' vectorial function F should be as consistent as possible with these information, that is to say, if two points (x_i and x_j) are measured as similar by S^a or S^b, they should receive similar classification vectors in F (F_i and F_j) and vice versa. On the other hand, if a data point x_i is within the initial label set, its classification vector F_i should be as consistent as possible with the initial labeling vector Y_i. In the following two subsections, we will present two different fusion schemes based on different optimization strategies, respectively.

3.2 Linear Fusion Scheme

In this scheme, the constraints from S^a, S^b and Y are fused simultaneously by a weighted sum. To meet this end, we formulate a regularized optimization framework by defining the following cost function with respect to F, which is a direct extension of [30]:

$$Q(F)=\left\{\mu\sum_{i,j=1}^{n}W_{ij}^{a}\left\|\frac{1}{\sqrt{D_{ii}^{a}}}\cdot F_i-\frac{1}{\sqrt{D_{jj}^{a}}}\cdot F_j\right\|^2+\right.$$
$$\left.\eta\sum_{i,j=1}^{n}W_{ij}^{b}\left\|\frac{1}{\sqrt{D_{ii}^{b}}}\cdot F_i-\frac{1}{\sqrt{D_{jj}^{b}}}\cdot F_j\right\|^2+\varepsilon\sum_{i=1}^{n}\left\|F_i-Y_i\right\|^2\right\} \quad (2)$$

The first, second and third items on the right hand side of Eq.2 correspond to the constraints from S^a, S^b and Y, respectively. The trade-off among these constraints is captured by the regularization parameters μ,η and ε, where $0<\mu,\eta,\varepsilon<1$ and $\mu+\eta+\varepsilon=1$.

Eq. 2 can be re-written in a more concise form as follows:

$$Q(F)=\mu F^T(I-S^a)F+\eta F^T(I-S^b)F+\varepsilon(F-Y)^T(F-Y) \quad (3)$$

With the above optimization criterion, the optimal vectorial function F^* is achieved when $Q(F)$ is minimized:

$$F^*=\arg\min_F Q(F) \quad (4)$$

Differentiating $Q(F)$ defined by Eq. 3 with respect to F leads to the following optimal vectorial function F^* (Linear Form):

$$F^*=(1-\mu-\eta)(I-\mu S^a-\eta S^b)^{-1}\cdot Y \quad (5)$$

Although the closed form is achieved, in some practical cases, the iterative form might be more preferable. We also develop an iterative version in Eq. 6:

$$F(t+1)=\mu S^a F(t)+\eta S^b F(t)+(1-\mu-\eta)Y$$
$$where\ F(0)=Y \quad (6)$$

By a similar analysis as in [30, 31], the relationship between closed form and iterative form can be given as:

$$F^*=\lim_{t\to\infty}F(t) \quad (7)$$

3.3 Sequential Fusion Scheme

Different from linear scheme, in this scheme, the constraints from S^a, S^b and Y are fused sequentially. In this case, we can formulate the following two-stage optimization problem:

$$Q_1(F)=\mu F^T(I-S^a)F+(1-\mu)(F-Y)^T(F-Y)$$
$$F_1^*=\arg\min_F Q_1(F) \quad (8)$$

$$Q_2(F)=\eta F^T(I-S^a)F+(1-\eta)(F-F_1^*)^T(F-F_1^*)$$
$$F_2^*=\arg\min_F Q_2(F) \quad (9)$$

Eq. 8 (Stage 1) defines an optimal F_1^* by considering the constraints from S^a and Y; while Eq. 9 (Stage 2) defines an optimal F_2^* by considering the constraints from S^b and F_1^*. The final classification or retrieval decision can be made based on F_2^*. The trade-off between S^a and Y is captured by the regularization parameters μ; while the trade-off between S^b and F_1^* is captured by the regularization parameter η, where $0<\mu,\eta<1$.

Solving the optimization problem defined by Eq. 8 and 9 leads to the following optimal vectorial function F_2^* (Sequential Form):

$$F_2^*=(1-\mu)(1-\eta)(I-\eta S^b)^{-1}(I-\mu S^a)^{-1}\cdot Y \quad (10)$$

Like in linear form, the iterative form for F_2^* can be given as :

$$F_2(t+1)=\mu S^a F_2(t)+\eta S^b F_2(t)-\mu\eta S^b S^a F_2(t)+(1-\mu)(1-\eta)Y$$
$$where\ F_2(0)=Y \quad (11)$$

And the relationship between the iterative form and linear form can be given as:

$$F_2^*=\lim_{t\to\infty}F_2(t) \quad (12)$$

3.4 Similarity Propagation

Using Taylor expansion and omitting the constant coefficient $(1-\mu-\eta)$ of Eq.5,:

$$F^*=(I-\mu S^a-\eta S^b)^{-1}\cdot Y$$
$$=\sum_{i=0}^{\infty}(\eta S^b+\mu S^a)^i Y \quad (13)$$
$$=Y+\left\{\eta S^b Y+\mu S^a Y\right\}+\left\{\eta S^b(\eta S^b Y+\mu S^a Y)+\mu S^a(\eta S^b Y+\mu S^a Y)\right\}+\cdots$$

Similarly, the optimal solution for sequential form as Eq. 10 can be re-written as:

$$F_2^*=\sum_{i=0}^{\infty}(\eta S^b)^i\sum_{j=0}^{\infty}(\mu S^a)^j Y$$
$$=\sum_{i=0}^{\infty}\sum_{j=0}^{\infty}(\eta S^b)^i(\mu S^a)^j Y \quad (14)$$

From the above equations, we can grasp the idea of the proposed method from a transductive learning point of view. Both F^* and F_2^* can be regarded as the sum of a series of infinite terms. The first term is simply the score of initial labels Y, the second term is to spread the scores of the initial labeled points to their nearby points by the two graphs; the third term is to further spread and fuse the scores, and so on. Thus the effect of un-labeled points is gradually incorporated; and the evidence from two modalities is fused at each step by weighted sum.

Comparing Eq. 13 and 14, it can be seen that the difference of the two proposed schemes lies in the ways they propagate and fuse the similarity by the two graphs. For linear form, the second term in Eq. 13 is to spread the scores of the initial labeled points to their nearby points by the two graphs S^a and S^b respectively, and then fuse the results by weighted sum; the third term is to further spread and fuse the scores, and so on. For sequential form, every item in Eq. 14 denotes 1) spreading initial label Y through S^a by arbitrary steps and then 2) spreading the result through S^b by arbitrary steps.

3.5 Bayesian Interpretation

Following [28], we present a Bayesian interpretation for the proposed schemes. In this way, we will reveal the prior fusion nature of the proposed schemes. Moreover, we will show that the difference between linear form and sequential form also comes from different way they fuse the prior.

Let $P(F)$ denotes the prior probability of F, and $P(Y|F)$ be the conditional probability of Y. Then the optimal F estimation can be given by MAP:

$$\begin{aligned} F^* &= \arg\max_F \{\log(P(Y|F)) + P(F)\} \\ &= \arg\min_F \{-\log(P(Y|F)) - P(F)\} \end{aligned} \tag{15}$$

As in [28], the conditional probability is assumed as:

$$P(Y|F) = \frac{1}{Z}\exp(-\varepsilon\|F-Y\|^2) \tag{16}$$

where Z is a normalization constant and ε is a scaling factor.

The prior $P(F)$ is also assumed as in [28], except that in multi-modality, we have two possible distributions: ($P_a(F)$ from modality a and $P_b(F)$ from modality b)

$$\begin{aligned} P_a(F) &= \frac{1}{Z}\exp\{-F^T(I-S^a)F\} \\ P_b(F) &= \frac{1}{Z}\exp\{-F^T(I-S^b)F\} \end{aligned} \tag{17}$$

By fusing $P_a(F)$ and $P_b(F)$ properly, we can obtain the final distribution of $P(F)$.

The first fusion strategy is the same as [8]. (For detailed discussion for this fusion assumption, refer to [8]):

$$\begin{aligned} P(F) &= \frac{1}{Z}[P_a(F)]^{\mu}[P_b(F)]^{\eta} \\ &= \frac{1}{Z}\exp\{-\mu F^T(I-S^a)F - \eta F^T(I-S^b)F\} \end{aligned} \tag{18}$$

Substituting Eq.16 and Eq. 18 into Eq. 15, will lead to exactly the same optimization criteria as Eq. 3, based on which linear form as Eq. 5 can be derived.

On the other hand, if we take the following fusion strategy as Eq. 19, and substitute Eq. 16 and Eq. 19 into Eq. 15, we will get the optimization criteria as Eq 20.

$$P(F) = \frac{1}{Z}\frac{\exp\{-\mu F^T(I-S^a)F - \eta F^T(I-S^b)F\}}{\exp\{-F^T\eta(I-S^b)\mu(I-S^a)F\}} \tag{19}$$

$$\begin{aligned} F^* = \arg\min_F \{&\mu F^T(I-S^a)F + \eta F^T(I-S^b)F \\ &- \eta\mu F^T(I-S^b)(I-S^a)F + \varepsilon\|F-Y\|^2\} \end{aligned} \tag{20}$$

Suppose $\mu+\eta-\eta\mu+\varepsilon=1$, the solution for the above equation is given as follows:

$$\begin{aligned} F^* &= \varepsilon(I-\mu S^a-\eta S^b+\eta\mu S^b S^a)^{-1}Y \\ &= (1-\mu)(1-\eta)(I-\eta S^b)(I-\mu S^a)^{-1}Y \end{aligned} \tag{21}$$

Note that this solution is exactly the same as the sequential form as in Eq. 10. Moreover, Eq. 20 can be viewed as the regularized optimization criteria for sequential form. In this way, sequential form actually can be optimized in one stage as linear form.

4. GRAPH BASED UN-SUPERVISED LEARNING IN MULTI-MODALITY

Define $S' = \frac{\mu S^a + \eta S^b}{\mu+\eta}$, and $\alpha' = \mu+\eta$ in Eq. 5, the closed form solution for linear form is converted to:

$$F^* = (1-\alpha')(I-\alpha' S')^{-1}Y \tag{22}$$

Similarly, if we define $S'_2 = \frac{\mu S^a + \eta S^b - \mu\eta S^b S^a}{\mu+\eta-\mu\eta}$ and $\alpha'_2 = \mu+\eta-\mu\eta$ in Eq. 15, the closed form solution for sequential form is converted to:

$$F_2^* = (1-\alpha'_2)(I-\alpha'_2 S'_2)^{-1}Y \tag{23}$$

Both Eq. 22 and 23 have exactly the same form as the original manifold ranking algorithm [30, 31]. In this way, the proposed algorithm can be viewed as an extension of manifold ranking algorithm in terms of new normalized similarity matrix learning.

Moreover, in [29], the graph Laplacian is defined as $\Delta = I - S$. Thus, the proposed algorithm can be also viewed as graph Laplacian fusion from multi-modality, with linear form:

$$\Delta' = I - \frac{\mu S^a + \eta S^b}{\mu+\eta} \tag{24}$$

and with sequential form as:

$$\Delta'_2 = I - \frac{\mu S^a + \eta S^b - \mu\eta S^b S^a}{\mu+\eta-\mu\eta} \tag{25}$$

Note that such similarity learning or Laplacian fusion is actually independent on the label information Y. Based on this observation, the proposed algorithm can be further extended to un-supervised learning by feeding the new normalized similarity matrix or graph Laplacian to some existing spectral methods.

In this paper, the spectral clustering algorithm in [16] is adopted for clustering task; while Laplacian Eigen-Map in [1] is used for embedding task. They are briefly summarized in algorithm 2 and 3, respectively.

Algorithm 2 Clustering from multi-modality

1. Define $S^{new} = \frac{\mu S^a + \eta S^b}{\mu+\eta}$ (linear form), or $S_{new} = \frac{\mu S^a + \eta S^b - \mu\eta S^b S^a}{\mu+\eta-\mu\eta}$ (sequential form);
2. Form the matrix X by staking the k largest eigen-vectors of S^{new}: $X = [x_1 x_2 \cdots x_k]$;
3. Normalize X so that each row of X has unit length;
4. Treat each row X of as a data point in R^k; and cluster them into k clusters using K-means;
5. Assign the original data point i to class j if and only if the i^{th} row of X belongs to cluster j

Algorithm 3 Embedding from multi-modality

1.	Define Ω^{new} by Eq. 24 (linear form), or Eq. 25 (sequential form);
2.	Form the matrix X by staking the k smallest eigen-vector of Ω^{new}: $X = [x_1 x_2 \cdots x_k]$;
3.	Normalize X so that each row of X has unit length;
4.	The embedding of original data point i in R^k is the i^{th} row of X.

Note that in algorithm 3, we made a modification to the optimization criteria of the original Eigen-Map algorithm [1] by:

$$\arg\min_{x_i} \{ x_i^T \Omega^{new} x_i \} \quad (i = 1, 2, \cdots, k) \tag{26}$$

5. EXPERIMENTAL RESULTS

5.1 Experimental Design

Three datasets are used in the experimental evaluation:

i) Web Page. This dataset is a subset of WebKB from [22]. 1051 Web pages are classified into two categories: 230 for 'course' and 821 for 'non-course'. Every Web page is represented by two modalities: plain text (modality a) and in-link anchor text (modality b). It has been used to evaluate the performance of Co-Train [3].

ii) Corel5000. This dataset consists of 5,000 Corel images. The images are categorized into 50 groups, each having 100 images. Images belonging to the same group are considered to be relevant. Every image is represented by two modalities: pyramid wavelet texture feature [15] (modality a) and color histogram [18] (modality b).

iii) Web Image. 9046 images are crawled from Web. Every image is manually annotated by 1 to 3 keywords from a pre-defined keyword list. There are totally 48 keywords in the keyword list, including 'Bear', 'Dinosaur', 'Orb', 'People', 'Flower', 'Shell' etc. Every image is represented by two modalities: the surrounding text[1] of the given image (modality a) and a combination of low-level feature as listed in table 1 (modality b).

Table 1. Feature combination for content modality

Feature Name	Dimension
color histogram [18]	36 Dim.
color correlogram [12]	144 Dim.
Tamura feature [21]	20 Dim.
pyramid wavelet texture feature [15]	24 Dim.

To construct the graph from image content modality (low-level feature for both Corel5000 and Web image), $L1$ distance is adopted to define the edge weights in the graph as Eq. 27, since it can better approximate the perceptual difference between two images than other popular Minkowski distances when using either color or texture representation or both [10]:

$$W_{ij} = \prod_{l=1}^{m} \exp\left(-\left| x_{il} - x_{jl} \right| / \sigma_l \right) \tag{27}$$

where x_{il} and x_{jl} are the lth dimension of x_i and x_j respectively; m is the dimensionality of the feature space; and σ_l is a positive parameter that reflects the scope of different dimensions and is set to 1 in this paper.

On the other hand, to construct the graph for text modality (surrounding text for Web image; plain text and in-link anchor text for Web page), the feature vector is weighted by TF/IDF [7] and the edge weight is defined by dot product:

$$W_{ij} = \sum_{l=1}^{m} x_{il} \cdot x_{jl} \tag{28}$$

where x_{il} and x_{jl} are the same as in Eq. 27 except that they represent text feature.

For all learning tasks, the proposed method (LIN for linear form and SEQ for sequential form) is compared with 1) using modality a only (AM), 2) using modality b only (BM), 3) using both modality a and b as one modality (AB-OM).

5.2 Experimental Results for Semi-supervised Learning

Both classification and image retrieval in the scenario of QBK are evaluated in this part. For both learning tasks, a small portion of data points are randomly selected from the dataset and manually labeled. In order to perform a systematic evaluation, we vary the size of training data, i.e. the number of initial manually labeled data points, and compare the average classification error or retrieval accuracy. Considering the randomness of the selection of initial labels, we run 10 times of labeling and training; and the average result is recorded.

Note that if sequential form is adopted, we need to determine the sequence of modality a and modality b (determining using $S^b S^a$ or $S^a S^b$ in Eq. 10). For simplification, we can substitute $S^b S^a$ in Eq. 10 by $\frac{1}{2}(S^b S^a + S^a S^b)$. However, in all of our experiments, we find that this sequence does not change the comparative results. Since our main concern in this paper is the comparative performance of the proposed schemas, we simply use $S^b S^a$ in Eq. 10 for all the datasets.

The regularized parameter for label information is fixed at 0.01, which is the same as in [10]. In this way, there is actually one remaining regularized parameter μ ($\mu + \eta = 0.99$ for linear form and $\mu + \eta - \mu\eta = 0.99$ for sequential form). A parametric study is conducted on μ using linear form. The range of μ to achieve satisfactory performance as well as the final value used in the experiments for different dataset is listed in table 2. In sequential form, we use the same value of μ as in linear form for the sake of simplicity.

[1] We use the same method as in [7] to extract the surrounding text of the given image.

Table 2. Parametric study on μ

Dataset	Satisfactory Range	Final Value
Web Page	$0.2 \le \mu \le 0.8$	$\mu = 0.5$
Corel 5000	$0.1 \le \mu \le 0.5$	$\mu = 0.3$
Web Image	$0.2 \le \mu \le 0.8$	$\mu = 0.5$

Dataset *i* (Web Page) is used to evaluate the classification performance. The classification error vs. training data curve is shown in Figure 1. The compared schemes use manifold ranking algorithm [30] for classification purpose. It can be seen from the figure that in most cases, using more feature does help to reduce the classification error. However, when training data size (the number of initial manually labeled data points) is 5, 65 or 100, AB-OM can not even beat AM or BM, indicating that using more feature as one modality actually causes performance deterioration in these cases. On the other hand, both proposed schemes outperform the other three in all cases. For example, when using 20 training data, the average classification errors for AB-OM, LIN, SEQ are 18.6%, 15.3%, 8.1%, respectively. Comparing LIN and SEQ, it can be seen that sequential form always outperforms linear form by a large margin.

Both Dataset *ii* and Dataset *iii* are used to evaluate the retrieval performance in the scenario of QBK. The average precision of top 20 retrieved images (P20) vs. training data size (the number of initial manually labeled data points) is shown in Figure 2. The average precision vs. scope is shown in Figure 3 when training data size is fixed at 100. The compared schemes use the algorithm in [20] to build the keyword model. For Corel5000, using two kinds of feature together always outperforms using only one of them in all cases. Comparing AB-OM, LIN and SEQ, it can be seen that treating all feature as two modalities (both LIN and SEQ) always outperform treating them as one modality (AB-OM), with SEQ achieving the highest accuracy.

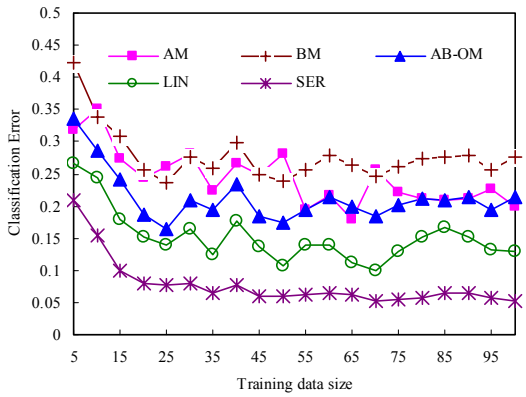
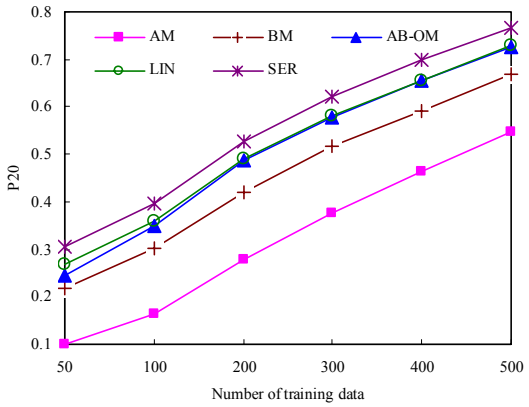
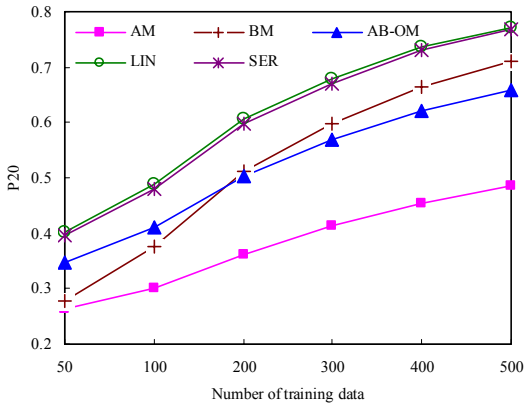


Figure 1. Systematic comparison of classification error under different sizes of training data



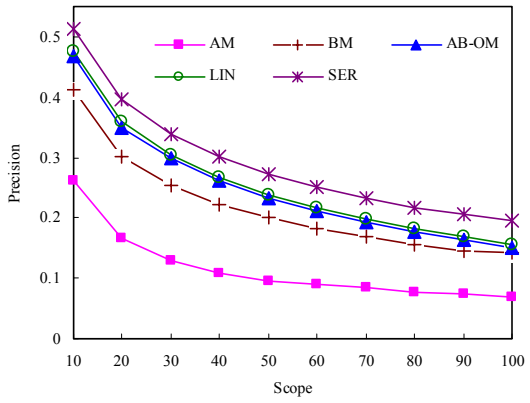
(a) Corel5000 dataset



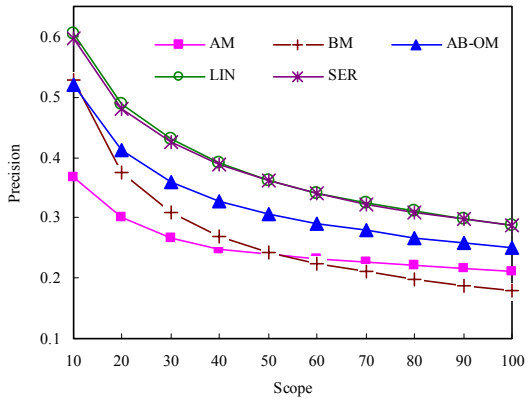
(b) Web Image dataset

Figure 2 P20 vs. training data size

For Web Image, it can be seen from Figure 2(b) that treating all feature as one modality (AB-OM) starts to cause performance degradation when training dataset size is greater than 200. In all cases, both the proposed schemes outperform the other three. Unlike in Corel5000, LIN is slightly better than SEQ for Web Image.



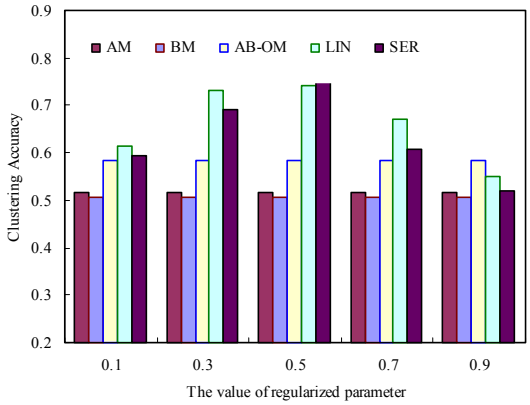
(a) Corel5000 dataset



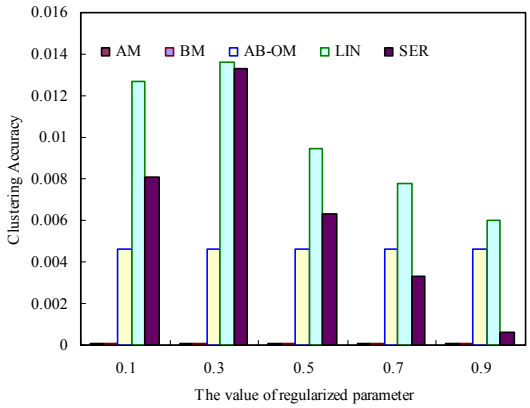
(b) Web Image dataset

Figure 3 Precision vs. scope when training data size is 100

(MI) are used for performance evaluation. (For detailed discussion of AC and MI, refer to [27].)



(a) Accuracy



(b) Normalized mutual information

Figure 4. Clustering result vs. regularized parameter μ

5.3 Experimental Results for Un-supervised Learning

Both clustering and embedding are evaluated in this part. When sequential form is adopted, the normalized similarity matrix or graph Laplacian might not be symmetrical. To address this issue, we add a pre-processing step in Algorithm 2 and 3:

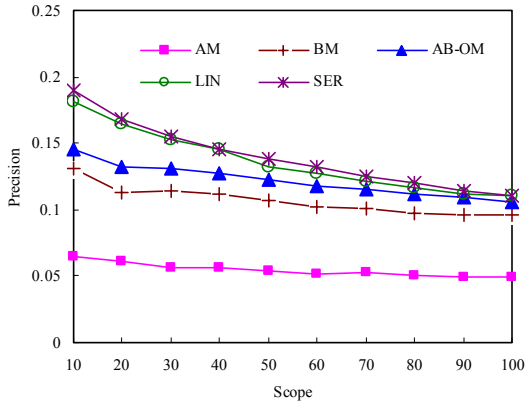
$$S^{new} \leftarrow \frac{1}{2}[S^{new} + (S^{new})^T]$$
$$\square^{new} \leftarrow \frac{1}{2}[\square^{new} + (\square^{new})^T] \quad (29)$$

2 classes are randomly selected form the Dataset *ii* dataset and fed into Algorithm 2. To perform a systematic evaluation, we run selecting and clustering 20 times and the average result is recorded. The compared schemes are fed into spectral clustering algorithm [16]. For all clustering algorithms, the reduced dimension (such as k in algorithm 2) is empirically set equal to the cluster number. As in [27], the accuracy (AC) and normalized mutual information

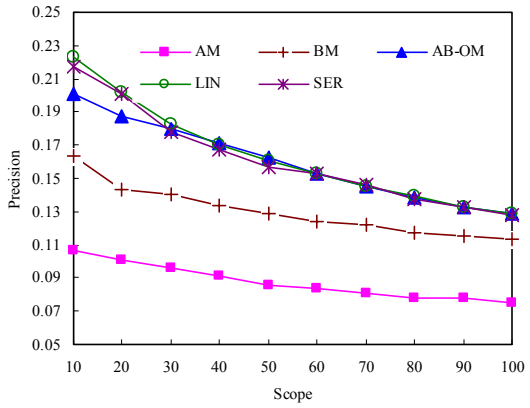
The results vs. the regularized parameter μ are shown in Figure 4. It can be seen from the figure that using two kinds of feature together always outperforms using only one of them in all cases. For accuracy, both LIN and SEQ outperform using two kinds of feature as one modality (AB-OM) when $0.1 \le \mu \le 0.7$. For normalized mutual information, while SEQ outperforms AB-OM when $0.1 \le \mu \le 0.5$, LIN remains best in all situations.

Dataset *ii* is used to evaluate embedding performance. The regularized parameter μ is set to 0.5 for both LIN and SEQ. After embedding by Algorithm 3, the image is indexed by its embedding. Then, we use each image in the database as a query, and average the results over the 5,000 queries. The precision vs. scope curve for different schemes when embedding dimension is set 10 and 100 is shown in Figure 6. Again, the proposed two schemes outperform the other three. For example, when the embedding dimension is fixed at 10, using only color feature or texture feature, P20 is 6.15% (AM) and 11.3% (BM), respectively; using color feature and texture feature as one modality, P20 is

13.2% (AB-OM); while P20 is 16.5% and 16.8% for LIN and SEQ, respectively.



(a) Embedding dimension is set 10



(b) Embedding dimension is set 50

Figure 5. Precision vs. scope for Corel5000 embedding

6. CONCLUSION

In this paper, we have made an investigation on graph-based learning methods in terms of their extension in multi-modality. For semi-supervised learning, two different fusion schemes, linear form and sequential form, are proposed. For each scheme, it is derived from optimization point of view and further justified from two sides: similarity propagation and Bayesian interpretation. By doing so, we reveal the regular optimization nature, transductive learning nature as well as prior fusion nature of the proposed schemes, respectively. Also, we show that the difference between the two schemes actually comes form 1) the different optimization strategy they adopt; 2) the different manner they spread and fuse similarity through graphs; and 3) the different way they fuse the prior probability. Moreover, the proposed method can be easily extended to unsupervised learning, including clustering and embedding. The effectiveness of the proposed method is justified by systematic experiments. In further work, we will 1) explore the working conditions of the two proposed schemes from a theoretical point of view; 2) investigate a more principled way to determine the regularized parameter; and 3) compare its performance with some existing multi-modality learning algorithms, such as Co-Train, super-kernel etc.

7. REFERENCES

[1] Belkin, M., and Niyogi, P. Laplacian Eigenmaps and spectral techniques for embedding and clustering. *Neural Computation,* pp. 1373-1396, 2003.

[2] Bickel, S., and Scheffer, T. Multi-view clustering. *Proc. of Int. Conf. on Data Mining*, pp. 19-26, 2004.

[3] Blum, A., and Mitchell, T. Combining labeled and unlabeled data with Co-Training. *Proc. of the Conf. on Computational Learning Theory*, pp. 92-100, 1998.

[4] Cai, D., He, X., Li, Z., Ma, W.Y., and Wen, J.R. Hierarchical clustering of WWW image search results using visual, textual and link information. *Proc. of the ACM Conf. on Information Retrieval*, pp. 952-959, 2004.

[5] Cascia, M.L., Sethi, S., and Sclaroff, S. Combining textural and visual cues for content-based image retrieval on the world wide web. *IEEE Workshop on Content-based Access of Image and Video Libaries*, pp. 24-28, 1998.

[6] Dupont, S., and Luettin, J. Audio-visual speech modeling for continuous speech recognition. *IEEE Trans. on Multimedia*, 2(3): 141-151, 2000.

[7] Feng, H., Shi, R., and Chua, T.S. A bootstrapping framework for annotating and retrieving WWW images. *Proc. of the ACM Int. Conf. on Multimedia*, pp. 960-967, 2004.

[8] Garg, A., Potamianos, G., Neti, C., and Huang, T.S. Frame-dependent multi-stream reliability indications for audio-visual speech recognition, *Proc. of Int. Conf. on Acoustics, Speech and Signal Processing*, vol. 1, pp. 24-27, 2003.

[9] Ghani, R. Combining labeled and unlabeled data multi-class text categorization. *Proc. of the Intl. Conf. on Machine Learning*, pp. 187-194, 2002.

[10] He, J., Li, M., Zhang, H.J., Tong, H., and Zhang, C. Manifold ranking based image retrieval. *Proc. of the ACM Conf. on Information Retrieval,* pp. 9-16, 2004.

[11] Heckmann, M., Berthommier, F., and Kroschel, K. Noise adaptive stream weighting in audio-visual speech recognition, *EURASIP Journal on Applied Signal Process*, pp. 1260-1273, 2002.

[12] Huang, J., Kumar, S.R., Mitra, M., Zhu, W.J., and Zabih, R. Image indexing using color correlograms. *Proc. IEEE Conf. on Computer Vision and Pattern Recognition,* pp. 762-768, 1997.

[13] Kailing, K., Kriegel, H., Pryakhin, A., and Schubert, M. Clustering multi-represented objects with noise. *Proc. of the Pacific-Asia Conf. on Knowledge Discovery and Data Mining*, pp. 394-403, 2004.

[14] Kittler, J., Hatef, M., and Duin, R.P.W. Combining classifiers. *Pattern Recognition*, pp. 897-901, 1996.

[15] Mallat, S.G., A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Trans. on Pattern Analysis and Machine Intelligence,* vol. 11, no. 7, pp. 674-693, 1989.

[16] Ng, A.Y., Jordan, M.I., and Weiss, Y. On spectral clustering:

analysis and an algorithm. *Advances in Neural Information Processing Systems,* 2001.

[17] Nigam, K., and Ghani, R. Analyzing the effectiveness and applicability of Co-Training. *Proc. of Information and Knowledge Management*, pp. 86-93, 2000

[18] Swain, M., and Ballard, D. Color indexing. *Int. Journal of Computer Vision*, 7(1): 11-32, 1991.

[19] Suen, C.Y., and Lam, L. Multiple classifier combination methodologies for different output level. *Proc. of the First Int. Workshop on Multiple Classifier*, pp. 52-66, 2000.

[20] Reference removed for double-blind review

[21] Tamura, H., Mori, S., and Yamawaki, T. Textural features corresponding to visual perception. *IEEE Trans. on Systems., Man and Cybernetics*, pp. 460-472, 1978.

[22] The WebKB dataset. http://meganesia.int.gu.edu.au/~phmartin/WebKB/.

[23] Wang, J., Zeng, H., Chen, Z., Lu, H., Tao, L., and Ma. W.Y. Recom: reinforcement clustering of multi-type interrelated data objects. *Proc. of the ACM Conf. on Information Retrieval*, pp. 274-281, 2003.

[24] Wu, Y., Chang, E.Y., Chang, K.C.C., and Smith, J.R. Optimal multimodal fusion for multimedia data analysis. *Proc. of the ACM Int. Conf. on Multimedia*, pp. 572-579, 2004.

[25] Yan, R., and Hauptmann, A.G. The combination limit in multimedia retrieval. *Proc. of the ACM Int. Conf. on Multimedia*, pp. 339-342, 2003.

[26] Yi, X. Zhang, C, and Wang, J. Multi-view EM algorithm and its application to color image segmentation. *IEEE Int. Conf. on Multimedia and Expo*, pp. 351-354, 2004.

[27] Zheng, X., Cai, D., He, X., Ma, W.Y., and Lin, X. Locality preserving clustering for image database. *Proc. of the ACM Conf. on Information Retrieval*, pp. 885-891, 2004.

[28] Zhou, D., and Schölkopf, B. A regularization framework for learning from graph data. *Workshop on Statistical Relational Learning at Int. Conf. on Machine Learning*, pp. 132-137, 2004.

[29] Zhou, D., and Schölkopf, B. *Transductive Inference with Graphs*. MPI Technical Report , 2004.

[30] Zhou, D., Bousquet, O., Lal, T.N., Weston, J., and Schölkopf, B. Learning with local and global consistency. *18th Annual Conf. on Neural Information Processing Systems*, pp. 237-244, 2003.

[31] Zhou, D., Bousquet, O., Lal, T.N., Weston, J., and Schölkopf, B. Ranking on data manifolds. *18th Annual Conf. on Neural Information Processing System*, pp. 169-176, 2003.