Nonnegative Shared Subspace Learning and Its Application to Social Media Retrieval

Sunil Kumar Gupta, Dinh Phung, Brett Adams, Truyen Tran and Svetha Venkatesh
Department of Computing
Curtin University of Technology
Perth, Western Australia
sunil.gupta@postgrad.curtin.edu.au,{d.phung,b.adams,t.tran2,s.venkatesh}@curtin.edu.au

ABSTRACT

Although tagging has become increasingly popular in online image and video sharing systems, tags are known to be noisy, ambiguous, incomplete and subjective. These factors can seriously affect the precision of a social tag-based web retrieval system. Therefore improving the precision performance of these social tag-based web retrieval systems has become an increasingly important research topic. To this end, we propose a shared subspace learning framework to leverage a secondary source to improve retrieval performance from a primary dataset. This is achieved by learning a shared subspace between the two sources under a joint Nonnegative Matrix Factorization in which the level of subspace sharing can be explicitly controlled. We derive an efficient algorithm for learning the factorization, analyze its complexity, and provide proof of convergence. We validate the framework on image and video retrieval tasks in which tags from the LabelMe dataset are used to improve image retrieval performance from a Flickr dataset and video retrieval performance from a YouTube dataset. This has implications for how to exploit and transfer knowledge from readily available auxiliary tagging resources to improve another social web retrieval system. Our shared subspace learning framework is applicable to a range of problems where one needs to exploit the strengths existing among multiple and heterogeneous datasets.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms

Algorithms, Theory

Keywords

nonnegative shared subspace learning, transfer learning, social media, image and video retrieval

KDD'10, July 25–28, 2010, Washington, DC, USA.

1. INTRODUCTION

Social tagging by users is a defining characteristic of Web 2.0 and has had a huge impact on the way we use the Web in a relatively short time. Social tagging systems exist that allow users to annotate and retrieve *any* Web-accessible item of interest, including web pages, images, sounds, videos, blog posts, tweets, links, URLs, locations, and even people. Similar to keywords in Information Retrieval (IR), short textual descriptors, termed tags, provide concise summarization of resources, often at topical or conceptual levels, which are difficult or impossible to infer using automatic, content-based IR methods. The resulting aggregation of tags forms a *folksonomy*, which acts as a proxy for a controlled taxonomy created by information science experts, and can be used to facilitate retrieval from the resources covered by the folksonomy. Folksonomy-enabled search has been instrumental in the rising popularity of social image and video sharing platforms, such as Flickr, Picassa and YouTube.

However, the use of tags poses serious challenges; The lack of constraints when creating free-text tags are part of their appeal, but as a result they tend to be noisy, ambiguous and incomplete [18, 14, 8], and could seriously degrade the retrieval performance. Research has attempted to improve the accuracy of tags [18, 24, 14, 27], but a common characteristic of the proposed solutions is a focus solely *within* the internal structure of a given tagging system. However, so long as only internal data is sought, these methods are less likely to be able to break the retrieval barriers caused by the uncertainty and noise inherent within the tags. Therefore, we offer in this paper an alternative approach that diverges from these methods. Our goal is to develop models to leverage external auxiliary sources of information to improve retrieval precision and recall in a target tagging system, presumably to be much noisier. The key intuition is that, by exploiting the common and disparate characteristics of a target domain with an appropriate auxiliary source, the retrieval performance in the target domain could be improved thanks to the reduction in the uncertainty of tags achieved through the auxiliary source. However, this gain is not always obvious – thus, it rises another important issue: what is the optimal level of joint modelling for which the target domain still benefits from the auxiliary source. To this end, we propose in this paper a shared subspace learning framework based on a joint nonnegative matrix factorization framework. The key advantage in our framework is the modeling flexibility to *explicitly* vary the level of joint modeling between data sources – and, as demonstrated experimentally, modeling optimal level of sharing results in significant improvement over existing method that perform joint matrix factorization without proper guidance (e.g. [27, 16]).

Specifically, our proposed shared subspace factorization learns co-occurrences from the subspaces of the target and auxiliary datasets by explicitly learning a common subset of basis vectors, but crucially number of shared basis vectors can be varied. This model also imposes the non-negativity constraint on the decomposed matrices, meaning that the basis vectors are part-based and hence represent the important, locally meaningful semantic features of the Web media. Significantly, the learnt subspace is a sparse representation of the data. NMF is also desirable for its ability to resolve the well-known 'polysemy' and 'synonymy' problems, known to exist in collaborative tagging applications which can cause performance degradation [8].

Our contributions are as follows. A novel formulation to exploit and transfer knowledge from auxiliary tagging resources to improve social image and video retrieval. We derive an efficient algorithm for learning the factorization, analyze its complexity, and provide proof of convergence. We provide evidence to validate our framework on image and video retrieval tasks from a Flickr and a YouTube dataset respectively using the LabelMe dataset to improve performance.

The novelty of our approach lies in the flexibility that permits the amount of subspace sharing to be varied from none to full sharing. Whilst these extremes have been considered before [13, 16], our results show that the best performance is achieved when both individual *and* joint subspaces are present together. Our work shares certain intuition with self-taught learning [20] and multi-view learning [9]. However, self-taught learning targets a supervised learning task using an auxiliary source of information, whereas we address an unsupervised learning task. Multi-view learning formulations, such as canonical correlation analysis (CCA) [9], learn two maximally correlated subspaces from two datasets without explicitly controlling the shared basis vectors and require two datasets such that each example in the first dataset correspond to one example in the second dataset and therefore can not be used in contexts when such one-to-one correspondences are not available (e.g. context of this paper). Our framework doesn't require any such one-to-one correspondences and hence provides potentially wider applicability.

The significance of the proposed shared subspace learning framework is firstly increased efficacy of image retrieval in Flickr and video retrieval in YouTube, Secondly, the framework has broader application to unsupervised learning with knowledge transfer from one domain to another of different quality. For example, the specific learning experiment we focus on can be perceived as transferring the view of LabelMe images provided by its folksonomy–static, visual objects, typified by nouns–to the Flickr images; But it might be just as desirable to transfer learning from an event or action-oriented folksonomy, typified by verbs, onto the LabelMe dataset. The Web continues to spawn new and specialized folksonomies, and the ability to cross-leverage these otherwise fragmented resources is desirable.

The rest of the paper is organized as follows. Section 2 briefly covers the necessary background for the paper. Section 3 presents the Joint Shared NMF (JSNMF) framework and describes the shared subspace learning. Section 4 describes the tag based social image/video retrieval using JSNMF. Section 5 presents the experimental results and conclusions are drawn in Section 6. Necessary proofs and derivations are pushed back to Appendix A and B.

2. BACKGROUND

From a broad perspective, previous work on tagging systems has been aimed at finding tag relevance, often improving tags by modifying them or recommending additional tags. Marlow et al. [18] present a taxonomy of social tagging systems, and highlight the effect of different design parameters and user communities on the make-up of the resulting tags. Folksonomies acquire differing characteristics by virtue of the myriad contextual factors and design decisions that lead to their creation. E.g., tagging systems that allow users to see the tags of others allow for rapid vocabulary convergence as opposed to "blind" systems; Users who tag solely for the purposes of retrieving their own resources are more likely to tag with idiosyncratic terms. Some folksonomies may be rich in conceptual labels, subsidiary descriptions etc., whilst others may consist predominantly of precise labels for visible objects.[1] They also present information on potential evaluative frameworks by providing a simple taxonomy of incentives and contribution models. Sigurbjörnsson and Zwol [24] investigate how users tag photos and the information contained in Flickr tags. Their analysis includes comment on the characteristics of tags in social web sites such as Flickr, and they provide a method to recommend a set of relevant tags at different levels of exhaustiveness from the original tags. Recent works by Li et al. [14, 15] present a method to learn social tag relevance by finding visual neighbors and voting based on tag frequency. The authors list all images for a given tag and then count the number of images which are visually similar to calculate tag relevance. Wang et al. [25] propose a search-based method which uses content-based retrieval technology to get visually similar images from an image collection and fuse it with text-based retrieval results. These methods perform poor in practice as visual similarity techniques have not yet matured. In addition, due to their need to search for visual neighbors, these methods are computationally expensive. In another work by Wu et al. in [27], authors note problems caused by irrelevant tags and describe the semantic loss caused when users do not tag a complete image but only one or two objects in the image. They propose a multi-modality tag recommendation method based on both tag and visual correlation by generating a ranking feature that uses each modality and therefore again suffer from problems faced by content-based techniques. Note that all the aforementioned approaches are confined to the noisy tags within the primary dataset of interest. Often, the quality of user-contributed tags is so poor that none of the above methods work well and an alternative method for improving retrieval performance must be sought.

One such alternative is to use auxiliary sources of information, resonating with the call in [11] which emphasizes the importance of the use of auxiliary source in the form of web-data and propose that the judicious use of such easily available data can substantially improve precision and recall. They also emphasize the need to formalize the notion of auxiliary sources of information in a rigorous framework. There exist many tagged datasets which are suitable for the purpose of enhancing performance of social media retrieval from a primary dataset, and are readily available. In particular, LabelMe [21] and Caltech-101 [7] are often used for evaluating and benchmarking object detection and classification techniques. In these datasets, users follow certain guidelines and\or a controlled vocabulary to construct groundtruth labels or tags, making them ideal auxiliary sources of information. To provide an assessment of how much sharing some of these datasets have with respect to tags, we provide, in Table 1, the Jensen-Shannon divergence between the tag distributions of different dataset pairs used in this paper.

On the technical side, our work falls in the field of transfer learning [19, 5] which deals with the transfer of knowledge across different domains (or tasks) which share some underlying structure.

[1] For example, a comparative analysis of the LabelMe and Flickr datasets reveals the Flickr tags to contain five times more function words (which exist only for grammatical purposes, and do not correspond directly to visual existents within an image) than LabelMe, and also contains less nouns.

Dataset Pairs	*Jensen-Shannon Divergence*
LabelMe-Flickr	0.5237
LabelMe-YouTube	0.5603
LabelMe-LabelMe	0.0758
YouTube-YouTube	0.1985
Flickr-Flickr	0.4511

Table 1: Jensen-Shannon Divergence between the tag distributions of different dataset pairs. For example, by "LabelMe-LabelMe", we mean two different subsets of LabelMe data .

In this setting, our proposed framework assumes that there is a common underlying subspace shared by the primary and secondary domains. However, unlike multi-task transfer learning approaches which focus on enhancing all the tasks, our approach focuses on only enhancing the performance of primary task using the secondary task. In particular, our work can be categorized as an instance of unsupervised transfer learning since there are no labels in either domains. Important difference from the previous unsupervised transfer learning approaches (see survey in [19]) is that our approach uses the representation of the data by not just using the shared subspace (common features) but also their private subspaces (individual features). Moreover, we are the first to present an unsupervised transfer learning framework for social tag based web retrieval task.

Our joint shared subspace learning method is formulated under the framework of nonnegative matrix factorization (NMF), a model widely used in text mining applications [23, 4, 3]. Formally, NMF aims to factorize a data matrix $\boldsymbol{X}$ into a product of a matrix $\mathbf{F}$ whose columns span the latent subspace and an encoding matrix $\mathbf{H}$:

$$\boldsymbol{X} \approx \mathbf{F}\mathbf{H}$$

where $\boldsymbol{X}$ is a $M \times N$ nonnegative data matrix containing N documents in terms of M vocabulary words , $\mathbf{F}$ ($M \times R$ nonnegative matrix) represents R basis vectors and $\mathbf{H}$ ($R \times N$ nonnegative matrix) contains the co-ordinates (if imagined in Euclidean space) of each document in the space spanned by the columns of the matrix $\mathbf{F}$. The part based nature of NMF comes from the nonnegativity constraint which is imposed on the matrices $\boldsymbol{X}$, $\mathbf{F}$ and $\mathbf{H}$ in the above factorization. Also, the decomposition achieves some level of sparsity [13] due to the nonnegativity of matrix $\mathbf{H}$ as the basis vectors (parts) can only be added and hence participate in a sparse manner to create the "whole". A special case of the proposed model is utilized for social and semantic analysis using NMF framework in Wu et al [28]. Another instance to discover temporal patterns in social media streams is proposed in [16].

3. NONNEGATIVE SHARED SUBSPACE LEARNING

We present a framework that captures the shared basis vectors between the two datasets and their individual bases corresponding to the discriminant subspace. This interpretation leads to the partitioning of the subspace into two parts. The first one is common to the datasets and the second is representative of the dataset in consideration. Let us represent the two datasets by $\boldsymbol{X}$,$\boldsymbol{Y}$ with dimension $M \times N_1$ and $M \times N_2$ respectively and write the decomposition and partition of the matrices in the following manner as :

$$\boldsymbol{X} \approx \underbrace{[\mathbf{W} \mid \boldsymbol{U}]}_{\mathbf{F}}\ \mathbf{H} = \mathbf{F}\mathbf{H} \tag{1}$$

$$\boldsymbol{Y} \approx \underbrace{[\mathbf{W} \mid \boldsymbol{V}]}_{\mathbf{G}}\ \mathbf{L} = \mathbf{G}\mathbf{L} \tag{2}$$

where $\mathbf{W}$ is a $M \times K$ matrix whose columns span the common subspace; $\boldsymbol{U}$ and $\boldsymbol{V}$ represent the remaining subspaces having dimension of $M \times (R_1 - K)$ and $M \times (R_2 - K)$respectively. K is the number of shared basis vectors and R_1 and R_2 are the dimensionality of low-rank underlying subspaces for $\boldsymbol{X}$ and $\boldsymbol{Y}$. R_1 and R_2 can be interpreted as the number of topics similar to the basic NMF case. $\mathbf{H}$ and $\mathbf{L}$ are the encoding matrices and have the dimension of $R_1 \times N_1$ and $R_2 \times N_2$ respectively and $\mathbf{F} \triangleq [\mathbf{W} \mid \boldsymbol{U}]$ and $\mathbf{G} \triangleq [\mathbf{W} \mid \boldsymbol{V}]$. Note that though usually $\boldsymbol{X}$ and $\boldsymbol{Y}$ have different vocabularies but they can be merged together to construct a common vocabulary that has M words.

We further impose a constraint of nonnegativity to achieve part-based representation. By nonnegativity, we mean that the elements of matrices $\mathbf{W}, \boldsymbol{U}, \boldsymbol{V}, \mathbf{H}$, and $\mathbf{L}$ are restricted to take only nonnegative values. To learn the required subspaces, we minimize the Frobenius norm of the joint decomposition error in the following manner:

$$\min_{\mathbf{W},\boldsymbol{U},\boldsymbol{V},\mathbf{H},\mathbf{L} \geq 0} \left\{ \frac{\|\boldsymbol{X} - [\mathbf{W} \mid \boldsymbol{U}]\mathbf{H}\|_F^2}{\|\boldsymbol{X}\|_F^2} + \frac{\|\boldsymbol{Y} - [\mathbf{W} \mid \boldsymbol{V}]\mathbf{L}\|_F^2}{\|\boldsymbol{Y}\|_F^2} \right\}$$

which can be translated into an minimizing problem with the following objective function

$$\min D \triangleq \frac{1}{2} \left\{ \|\boldsymbol{X} - [\mathbf{W} \mid \boldsymbol{U}]\mathbf{H}\|_F^2 + \lambda \|\boldsymbol{Y} - [\mathbf{W} \mid \boldsymbol{V}]\mathbf{L}\|_F^2 \right\}$$
$$\text{subject to } \mathbf{W}, \boldsymbol{U}, \boldsymbol{V}, \mathbf{H}, \mathbf{L} \geq 0$$

where $\|.\|_F$ is the Frobenius norm and $\lambda = \|\boldsymbol{X}\|_F^2 / \|\boldsymbol{Y}\|_F^2$ is defined to be the relative ratio between Frobenius norms of the two data matrices.

Expressing D elementwise, this optimization can be efficiently solved in similar fashion to the original formulation of NMF [13], yielding the following multiplicative update equations for $\mathbf{W}$ (see the appendix for the detailed proof):

$$(\mathbf{W})_{ab} \leftarrow (\mathbf{W})_{ab} \times (\mathbf{S})_{ab} \tag{3}$$

where $(\mathbf{S})_{ab}$ is given by

$$1/(\mathbf{S})_{ab} = \frac{\left(\mathbf{W}\mathbf{H}_w\mathbf{H}_w^{\mathsf{T}} + \boldsymbol{U}\mathbf{H}_u\mathbf{H}_w^{\mathsf{T}}\right)_{ab}}{\left(\boldsymbol{X}\mathbf{H}_w^{\mathsf{T}} + \lambda\boldsymbol{Y}\mathbf{L}_w^{\mathsf{T}}\right)_{ab}} + \lambda \frac{\left(\mathbf{W}\mathbf{L}_w\mathbf{L}_w^{\mathsf{T}} + \boldsymbol{V}\mathbf{L}_v\mathbf{L}_w^{\mathsf{T}}\right)_{ab}}{\left(\boldsymbol{X}\mathbf{H}_w^{\mathsf{T}} + \lambda\boldsymbol{Y}\mathbf{L}_w^{\mathsf{T}}\right)_{ab}}$$

and $\mathbf{H} \triangleq \left[\mathbf{H}_w^{\mathsf{T}} \mid \mathbf{H}_u^{\mathsf{T}}\right]^{\mathsf{T}}$and $\mathbf{L} \triangleq \left[\mathbf{L}_w^{\mathsf{T}} \mid \mathbf{L}_v^{\mathsf{T}}\right]^{\mathsf{T}}$.

Similar multiplicative update equations are obtained for $\boldsymbol{U}, \boldsymbol{V}, \mathbf{H}$ and $\mathbf{L}$:

$$(\mathbf{H})_{ab} \leftarrow (\mathbf{H})_{ab} \frac{\left(\mathbf{F}^{\mathsf{T}}\boldsymbol{X}\right)_{ab}}{\left(\mathbf{F}^{\mathsf{T}}\mathbf{F}\mathbf{H}\right)_{ab}} \tag{4}$$

$$(\mathbf{L})_{ab} \leftarrow (\mathbf{L})_{ab} \frac{\left(\mathbf{G}^{\mathsf{T}}\boldsymbol{Y}\right)_{ab}}{\left(\mathbf{G}^{\mathsf{T}}\mathbf{G}\mathbf{L}\right)_{ab}} \tag{5}$$

$$(\boldsymbol{U})_{ab} \leftarrow (\boldsymbol{U})_{ab} \frac{\left(\boldsymbol{X}\mathbf{H}_u^{\mathsf{T}}\right)_{ab}}{\left(\mathbf{W}\mathbf{H}_w\mathbf{H}_u^{\mathsf{T}} + \boldsymbol{U}\mathbf{H}_u\mathbf{H}_u^{\mathsf{T}}\right)_{ab}} \tag{6}$$

$$(\boldsymbol{V})_{ab} \leftarrow (\boldsymbol{V})_{ab} \frac{\left(\boldsymbol{Y}\mathbf{L}_v^{\mathsf{T}}\right)_{ab}}{\left(\mathbf{W}\mathbf{L}_w\mathbf{L}_v^{\mathsf{T}} + \boldsymbol{V}\mathbf{L}_v\mathbf{L}_v^{\mathsf{T}}\right)_{ab}} \tag{7}$$

These multiplicative update equations[2] obtained in our joint subspace learning case carry a similar intuition as in the NMF: if the perfect factorization is achieved, the multiplicative factors in the update equations reduce to unity. That is, it can be verified by inspection that if the factorization for the two data sets $\boldsymbol{X}$ and $\boldsymbol{Y}$ in equations (1) and (2) are exact, then the multiplicatives on the RHS of the update equations from (3) to (7) are unity. A pseudo code for our proposed nonnegative joint subspace factorization is shown in Algorithm 1. The convergence of this algorithm can also be proved with details given in the Appendix B.

Algorithm 1 Joint Shared Nonnegative Matrix Factorization (JSNMF).

```
Input: Datasets $\boldsymbol{X}, \boldsymbol{Y}$, Parameters $R_1$, $R_2$, $K$ and a threshold $\epsilon$
let $\lambda = \|\boldsymbol{X}\|_F^2 / \|\boldsymbol{Y}\|_F^2$
initialize $\mathbf{W}_0, \boldsymbol{U}_0, \boldsymbol{V}_0, \mathbf{H}_0, \mathbf{L}_0$ randomly
set $r = 1$
while ($r <$MaxNumIters) or ($C < \epsilon$) do
    update $\mathbf{W}_r, \boldsymbol{U}_r, \boldsymbol{V}_r, \mathbf{H}_r, \mathbf{L}_r$ according to eqs (3)– (7)
    normalize each column of $\mathbf{W}_r$, $\boldsymbol{U}_r$ and $\boldsymbol{V}_r$ to 1.
    let $\boldsymbol{X}_r = [\mathbf{W}_r \mid \boldsymbol{U}_r]\mathbf{H}_r$ and $\boldsymbol{Y}_r = [\mathbf{W}_r \mid \boldsymbol{V}_r]\mathbf{L}_r$
    compute error $C = \|\boldsymbol{X} - \boldsymbol{X}_r\|_F^2 + \lambda \|\boldsymbol{Y} - \boldsymbol{Y}_r\|_F^2$
    $r = r + 1$
end while
Output: return $\mathbf{W}, \boldsymbol{U}, \boldsymbol{V}, \mathbf{H}, \mathbf{L}$
```

We note two special cases from our joint factorization framework. When there is no sharing (i.e., $K = 0$ or $\mathbf{W}$ does not exist, and hence no common basis vectors between the two subspaces), the update equations for $\boldsymbol{U}$, $\boldsymbol{V}$, $\mathbf{H}$ and $\mathbf{L}$ reduce to individual NMF for $\boldsymbol{X}$ and $\boldsymbol{Y}$ described in [13]. Similarly, when we force a single shared subspace for the two datasets i.e., $\dim(\boldsymbol{U}) = 0$ and $\dim(\boldsymbol{V}) = 0$, the update equations reduce the fully joint formulation (recently studied in [28]).

For complexity analysis, we compare the cost of the basic NMF algorithm in [13] and our proposed JSNMF. For an $M \times N_1$ matrix $\boldsymbol{X}$ and an $M \times N_2$ matrix $\boldsymbol{Y}$, assuming that K basis vectors are shared and the latent space dimensionality for decomposition of $\boldsymbol{X}$ is R_1 and that of $\boldsymbol{Y}$ is R_2, then computational complexity for the JSNMF per iteration is $O(\max\{MN_1R_1, MN_2R_2\})$. The basic NMF algorithm in [13] applied for $\boldsymbol{X}$ and $\boldsymbol{Y}$ separately will have complexity of $O(MN_1R_1)$ and $O(MN_2R_2)$ respectively. This shows that JSNMF enjoys the same complexity as the basic NMF.

As far as the shared subspace dimensionality (K) is concerned, there does not seem to be any straight forward way to determine it exact value. However, empirical 'rule of thumb' methods have been used to a good degree of success. In our case, the value of K is bounded between 0 and $\min(R_1, R_2)$ i.e. ranges from no sharing to full sharing. Its optimal value depends on nature of the target and auxiliary data. Intuitively, K increases with the level of sharing between the two data sources. For an rough estimate on K, we find the number of the common features (tags in our case) between the two datasets, say M_{xy}, then the rule of thumb is to use $K = \sqrt{M_{xy}/2}$ as suggested by Mardia et al in [17]. Another way to estimate K is based on rank-estimation and as follows. Supposedly if subspaces spanned by $\mathbf{W}$, $\boldsymbol{U}$ and $\boldsymbol{V}$ are mutually-orthogonal then $K = rank\left(\boldsymbol{X}^\mathsf{T}\boldsymbol{Y}\right)$. In our case, however, $\mathbf{W}$, $\boldsymbol{U}$ and $\boldsymbol{V}$ are only approximately mutually-orthogonal, suggesting that optimal K can be approximated by $rank\left(\boldsymbol{X}^\mathsf{T}\boldsymbol{Y}\right)$. Intuitively, projection of $\boldsymbol{X}$ on $\boldsymbol{Y}$ implies the sharing level and hence determines K. However, this approach is computationally expensive.

4. SOCIAL IMAGE/VIDEO RETRIEVAL

In the light of JSNMF algorithm presented in previous section, the matrix $\boldsymbol{X}$ is taken as the *tf-idf* weighted [22] term-document matrix generated from the tags of target dataset and the matrix $\boldsymbol{Y}$ as the equivalent matrix generated from the tags of auxiliary dataset. We learn the joint shared subspace $\mathbf{W}$, the discriminant subspace $\boldsymbol{U}$ for target dataset, co-ordinate matrix $\mathbf{H}$ for target dataset, the discriminant subspace $\boldsymbol{V}$ for auxiliary dataset and co-ordinate matrix $\mathbf{L}$ for auxiliary dataset. Given a query sentence S_Q, a query vector q_x is constructed by expanding the query so that it includes all the unigrams (from vocabulary) which contain the words from the specified query sentence S_Q. The vector q_x is projected on the column space of matrix $[\mathbf{W} \mid \boldsymbol{U}]$ to find out the query encoding vector (let us denote by q_h). We compute the cosine similarity between q_h and the columns of matrix $\mathbf{H}$ to find out the similarly tagged images (or videos), and the results are ranked based on these similarity scores.

More formally, let the image (or video) dataset on which retrieval is to be performed, be represented as $\mathcal{I} = \{I_1, I_2, ..., I_{N_1}\}$ and the collection of associated tags as $\mathcal{T}$. We prepare term document matrix $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_{N_1}]$ where $\boldsymbol{x}_k$ $(k = 1, ...N_1)$ is *tf-idf* weighted term-document vector prepared using the tags of image (or video) I_k. We decompose matrix $\boldsymbol{X}$ to generate the matrices $\mathbf{W}$, $\boldsymbol{U}$ and $\mathbf{H}$ using the joint shared NMF technique described in section 3. Algorithm 2 provides pseudo-code for the social image/video retrieval using the proposed shared subspace learning framework where $\bullet$ and $\oslash$ denote element-wise matrix multiplication and division respectively.

Algorithm 2 Image/Video Retrieval using JSNMF.

```
Input: Given $\mathbf{W}, \boldsymbol{U}, \mathbf{H}$ (learnt using Algorithm 1), query sentence $S_Q$, number of images (or videos) to be retrieved $N$ and image (or video) dataset $\mathcal{I} = \{I_1, I_2, ......., I_{N_1}\}$
prepare $q_x$ using tf-idf method from $S_Q$
set $\delta = 10^{-9}$, $\epsilon = 10^{-2}$, $\mathbf{F} \triangleq [\mathbf{W} \mid \boldsymbol{U}]$ and project $q_x$ onto $\mathbf{F}$ to get $q_h$ by an initialization then looping as below
while $\left(\|\mathbf{F}q_h - q_x\|_2 \geq \epsilon\right)$ do
    $q_h^{r+1} \longleftarrow q_h^r \bullet \left(\mathbf{F}^\mathsf{T} q_x\right) \oslash \left(\mathbf{F}^\mathsf{T}\mathbf{F} q_h^r + \delta\right)$
end while
compute cosine similarities $\text{sim}(q_h, h_i)$ between query sentence and each tag document (image or video tags) according to the following
    $$\text{sim}(q_h, h_i) = \frac{q_h^\mathsf{T} h_i}{\|q_h\|_2 \|h_i\|_2}$$
sort the cosine similarities $\text{sim}(q_h, h_i)$ in descending order and get the ranking indices set $\{s_1, s_2, \ldots, s_{N_1}\}$
Output: return the top $N$ retrieved images (or videos) as $\{I_{s_k} \mid k = 1, ..., N\}$
```

5. EXPERIMENTS

5.1 Experimental Setup

Our experiment is setup to utilize tags from LabelMe (auxiliary) to improve two social web retrieval tasks (target) in two media domains : image (Flickr) and video (YouTube). We denote by $\boldsymbol{X}$, the

[2] Note that in the implementation, a common practice is to add a small number δ (we used $\delta = 10^{-9}$) in the denominator to avoid division by zero.

target dataset from which retrieval is to be performed and by $\boldsymbol{Y}$, the auxiliary data source (LabelMe dataset). Using the algorithm described in section 3, we learn a jointly shared subspace $\mathbf{W}$ using the two datasets and the discriminant subspaces $\boldsymbol{U}$ and $\boldsymbol{V}$ using the proposed JSNMF.

For comparison, we consider two baseline performances corresponding to two existing methods. In the first case, Flickr (or YouTube) data is used alone to learn NMF latent subspace $\boldsymbol{U}$ using the NMF algorithm [13]. In the second baseline, we consider a recent method proposed for social semantic analysis temporal patterns discovery in social media streams, which can also be utilized for image retrieval problem [16]. This method forces whole latent subspace to be shared between the two dataset, i.e., $\boldsymbol{U}$ and $\boldsymbol{V}$ vanishes. We shall call these two baselines $BaselineI$ and $BaselineII$ respectively. We note that in these two cases, it is likely that data faithfullness in each individual domain is not preserved. This is where the attractiveness of our framework lies: it provides a freedom to exploit as much sharing information as possible, but at the same time, respect the individual differences in each domain. Our experiment is centrally designed to evaluate this point in retrieval tasks. In the end, we also compare our best results with contemporary state-of-the-art techniques [15, 25] on image retrieval using Flickr dataset and present the comparison in Table 2.

We evaluate the proposed JSNMF framework against the two baseline methods outlined above in regard the two aspects (1) the effect of different level of sharing on the retrieval performance (2) and improvement in performance when auxiliary source of information is used with JSNMF compared to the use of tags from the same data source.

5.1.1 Data Collection and Groundtruth

Since there is no standard groundtruth data available to evaluate the proposed tasks, we construct a subset of images/videos crawled from Flickr, YouTube and LabelMe and manually evaluate the retrieval results. To obtain the data, we designed a set of concepts varying from indoor (e.g., 'chair', 'computer', 'cup', 'door', 'desk', 'microwave') to outdoor (e.g., 'beach', 'boat', 'building', 'plane', 'ship', 'sky', 'tree') and generic ('book', 'car', 'pen', 'person', 'phone', 'picture', 'window').

Flickr Dataset. We downloaded 50000 images from Flickr website using its APIs [2]. On average, the number of distinct tags are 8. We removed the rare tags (appearing less than 5 times in the entire corpus), images with no tags and the images having non-English tags. After the cleaning process, we obtained around 20,000 labeled images. From this data, 7000 examples are kept aside to be used as an auxiliary dataset (needed to investigate the use of auxiliary data from within domain).

YouTube Dataset. We downloaded 18000 videos' metadata (including tags, URL, category, title, comments etc) using YouTube API service [1]. On average, YouTube folksonomy has 7 distinct tags per video. As above, we remove the rare tags (appearing less than 2 times in the entire corpus), videos with no tags and non-English tags. After the cleaning process, we obtain around a dataset corresponding to 12000 videos. Again, we keep aside 7000 examples to be used as an auxiliary dataset (needed to investigate the use of auxiliary data from within domain).

LabelMe Dataset. Further we add around 7000 images with its tags from LabelMe. On average, there are 32 distinct tags per image. We removed the rare tags appearing less than 2 times in the entire data corpus. This process does not reduce the size of dataset.

5.2 Evaluation Measures

For the purpose of evaluation, we defined a query set $\mathbb{Q}$ = {'cloud', 'man', 'street', 'water', 'road', 'leg', 'table', 'plant', 'girl', 'drawer', 'lamp', 'bed', 'cable', 'bus', 'pole', 'laptop', 'plate', 'kitchen', 'river', 'pool', 'flower'}. Again, since there is no benchmark dataset available for evaluation, we construct the ground truth by manually going through each example in the two datasets (Flickr and YouTube) and annotating them with respect to this query set. We consider a query term and an image (or video) relevant if the concept is clearly visible in the image (or video).

To evaluate the overall retrieval performance, we use popular 11-point Average Precision-Recall Curve. For the web retrieval task, typically web users would like every item (images or videos) to be highly relevant with respect to the query in the first few retrieved results. Therefore, we also present the well-known *precision-scope (P@N)* curve[3] to clearly demonstrate the ranking of the relevant images (or videos) in the retrieved set. This measure is not calculated for the entire retrieved set. By looking at the retrieval result at various scope level, it is easier to appreciate the ranking performance of the method.

5.3 Flickr Image Retrieval Results

To compare the retrieval performance at different levels of sharing K, we consider the popular precision-at-fixed-recall metric in information retrieval. We fix the recall at 0.1, which is adequate since most of the time users are interested in only first few results. For the subspace learning, we set $R_1 = 60$ and $R_2 = 40$ respectively, interpreted as the latent dimensionalities in the data. Figure 1 presents the precision figures for the query set $\mathbb{Q}$ defined in subsection 5.2 in increasing values of the sharing dimension K.

As can be seen, the average precision follows an interesting bell-shape curve when K increases, suggesting that there is an optimal level of sharing to achieve the best level transfer from one data source to another. It also indicates that the two baselines with no or total sharing are highly non-optimal approaches. On an average across the query set $\mathbb{Q}$, $K = 15$ results in 58% precision and is the optimal level of sharing for our dataset. This contrasts with $50\%(K = 0, BaselineI)$ and $46\%(K = 40, BaselineII)$ and thus our method delivers around 10% improvement.

One might explain these results as follows: when K is much smaller than 15, the learnt subspace shares very few basis vectors with LabelMe dataset and therefore does not benefit fully by its accurately tagged nature - this is caused by an *under-representation* between two datasets. On the other hand, when K becomes much larger than 15, it forces many basis vectors in the learnt subspace to represent both the datasets which can be difficult and therefore, the approximation in JSNMF factorization becomes poor - and this is caused by an *over-representation* between two datasets. Neither extreme is desirable. The optimal sharing $K = 15$ in our framework represents a case in which an appropriate level of representation for the two datasets is achieved. To further highlight this effect, the retrieval performance in terms of average precision for different values of shared subspace dimensionality K is plotted in Figure 1

[3] In this curve N represents the number of top retrieved images with which we compute the precision. For example $P@20$ is the retrieval precision when considering only the first 20 images retrieved.

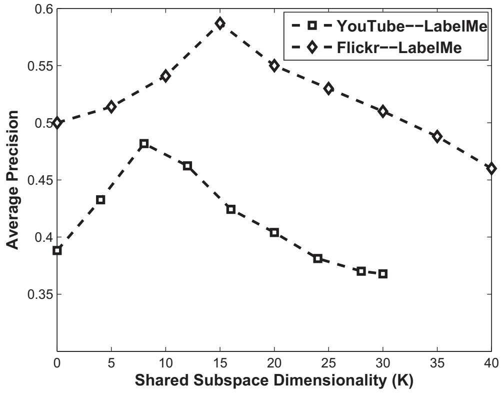


Figure 1: Retrieval performance with respect to shared subspace dimensionality for Flickr and YouTube.

where one can observe the bell-shape behaviors. We also note the correlation between the number of common basis vectors with the Jensen-Shannon divergence values given in Table 1.

To demonstrate the ranking capabilities of the JSNMF based retrieval, we compute the precision at various scope levels. Figure 2a depicts the retrieval performance in terms of average precision ($P@N$) with respect to scope values and MAP metrics and it is evident from the graph that JSNMF with $K = 15$ achieves much better performance than both $BaselineI$ and $BaselineII$.

Precision-Recall Curve

To examine the overall performance across all recall values, we present the standard 11-point average precision-recall curve for $K = 15$. To compare the performance against the two baselines, we also present these curves for $BaselineI$ and $BaselineII$ averaged over all queries in $\mathbb{Q}$. Figure 2b depicts these precision-recall curves. That shows that across all recall values, the JSNMF framework consistently demonstrates its benefits against the two baseline methods.

5.4 YouTube Video Retrieval Results

To demonstrate the flexibility of applying the framework, we conducted YouTube experiments in the same way as for the Flickr dataset. Again we fix the recall at 0.1 (due to users' interest in only first few results) and generated results at various levels of subspace sharing by varying K. This time the latent dimensionality for YouTube (R_1) was set to be 30 and that of LabelMe (R_2) was set to be 40 as before. Figure 1 presents the precision figures for the query set $\mathbb{Q}$ by increasing the sharing dimensionality (K) at a step of 4.

Similar to the Flickr case, the average precision follows a curve indicating an optimal level of sharing corresponding to $K = 8$. Using this optimal level of sharing ($K = 8$) between YouTube and LabelMe, we achieve improvement in precision performance by around 10% compared to $BaselineI$ and by around 12% compared to $BaselineII$. This clearly repeats the *under-representation and over-representation* observations made above for Flickr case.

To demonstrate the ranking capability along with overall performance, we, once again, present the Precision-Scope (P@N) and MAP results shown in Figure 3a. Note that the best performance in terms of both metrics has been achieved at the optimum sharing level ($K = 8$) and this result is consistently better than the two baselines.

Precision-Recall Curve

To evaluate the overall performance across all recall values, we present the standard 11-point average precision-recall curve for $K = 8$ found above to be optimal sharing level. To compare the performance against the two baselines, we also present these curves for $BaselineI$ and $BaselineII$ averaged over all queries in $\mathbb{Q}$. Figure 3b depicts these precision-recall curves. Again this figure shows that across all recall values, JSNMF method consistently demonstrates its benefits against the two baseline methods.

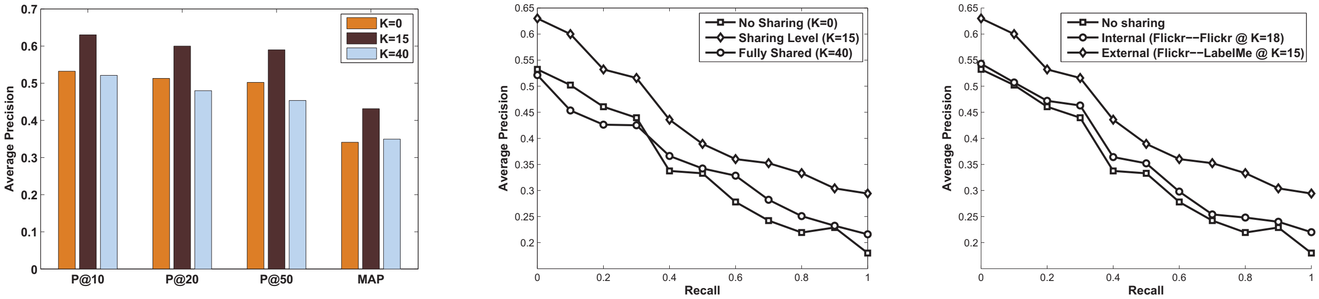
External vs. Internal Auxiliary Source

One might argue on the usefulness of auxiliary data; specifically, what if more data from within the internal system is used instead of the use of external data. Addressing this question, we investigate the benefits of LabelMe (external auxiliary source) vis-à-vis the tags from the same domain (internal auxiliary source), we denote YouTube dataset as $\boldsymbol{X}$ and another YouTube dataset as $\boldsymbol{Y}$. Then JSNMF algorithm is used to learn both $\mathbf{W}$ and $\boldsymbol{U}$. We repeat the experiment as conducted for YouTube and LabelMe sharing case and find that the optimal sharing in this turns out to be $K = 16$. Figure 3c clearly shows that improvement due to LabelMe data is much better than that achieved by internal auxiliary data (the second YouTube dataset). We believe that this improvement is due to the controlled, more complete and objective nature of LabelMe tags which helps in discovering the right term co-occurrences. Similar result is shown in Figure 2c for Flickr dataset where we compare the retrieval performance of Flickr dataset by sharing with another Flickr dataset (optimum sharing achieved at $K = 18$) and with LabelMe dataset (external). Again, the improvement due to LabelMe over noisy auxiliary Flickr data (the second Flickr dataset) is significant.

Finally, Table 2 presents the comparison of our results with the two contemporary works done [15, 25] on image annotation (for image retrieval) for Flickr dataset. This comparison is based on the precision figures presented by authors in [15]. Though the dataset used by the authors in [15, 25] is not identical but it is similar to us in the sense that both of these works use Flickr data and dataset size is of the same order. Instead of comparing the results at various parameter values, we compare the best results achieved by all three methods. In our work, we got the best results on Flickr dataset by sharing the subspace with LabelMe dataset at $K = 15$ which is significantly better than the results presented by the authors in [15] for 200 textual neighbors and 200 visual neighbors.

5.5 Discussion

Our proposed joint subspace learning framework requires the specification of the sizes of the latent dimensionalities R_1 and R_2 from the two datasets in the same way a standard NMF [13] requires the latent dimensionality for its subspace. This suggests that further regularization techniques such as analysis on sparsity as in [10] can be analyzed for our JSNMF. The additional requirement of K will inevitably incur some extra modeling in the regularization process, but still appears possible to be carried out in the same way as in [10]. Another limitation in the current setting of the JSNMF in this paper is that we limit our analysis to only two datasets. Extension to multiple datasets is also possible. Two apparent approaches are either to impose a common shared subspace $\mathbf{W}$ among all datasets, or perform a pairwise shared subspace learning $\mathbf{W}_{i,j}$ for each pair of the datasets. Both of these suggestions are feasible within the framework provided in this pa-



(a)

(b)

(c)

Figure 2: Flickr image retrieval results (a) Precision-Scope and MAP plots for $BaselineI$ ($K = 0$), optimally shared subspace ($K = 15$) and $BaselineII$ ($K = 40$) (b) 11-Point interpolated Precision-Recall Curve for $BaselineI$ ($K = 0$), optimally shared subspace ($K = 15$) and $BaselineII$ ($K = 40$) (c) Comparison with case when auxiliary data is chosen from within Flickr domain instead of LabelMe.

per. Alternatively, one might consider a more advanced statistical modeling and regularization among the subspaces of several variables [12], or probabilistic graphical modelling [26].

Recall Values	*Precision (approx.) Wang et al. [25] (Text-Neighbor200)*	*Precision (approx.) Li et al. [15] (Visual-Neighbor200)*	*Precision our method (JSNMF @ $K = 15$)*
0.1	0.27	0.29	0.60
0.2	0.18	0.22	0.54
0.3	0.10	0.13	0.51
0.4	0.08	0.11	0.44
0.5	0.05	0.07	0.38

Table 2: Comparison of our results with contemporary state-of-the arts on Flickr dataset.

6. CONCLUSION

We have presented a novel nonnegative shared subspace learning framework and applied it to improve tag-based image and video retrieval in online social image tagging systems (Flickr) and video sharing system (YouTube) respectively by leveraging an auxiliary source of information (LabelMe). Apart from possessing the same features as a typical NMF approach to text analysis, such as the ability to capture tag co-occurrences and part-based decomposition, a key feature of our proposed JSNMF is the ability to discover the shared structures between the two datasets and the flexibility in controlling the optimal level of sharing between them. This feature is important in dealing with real-world datasets since the practice of forcing the subspaces to be identical or totally different as done in current existing works is highly unrealistic. Our experimental results have consistently validated this point, showing that an appropriate level of subspace sharing can significantly boost the retrieval performance - an average of over 10% for retrieval precision on the Flickr dataset and an average of over 11% for retrieval precision on the YouTube dataset using LabelMe as the auxiliary source. Although applied to image and video retrieval in this paper, our shared subspace learning framework is generic and can be applied to a wider setting in machine learning and data mining tasks.

7. REFERENCES

[1] http://code.google.com/apis/youtube/overview.html. Accessed in Oct, 2009.

[2] http://www.flickr.com/services/api/. Accessed in July, 2009.

[3] H.D. Abdulla, M. Polovincak, and V. Snasel. Search results clustering using nonnegative matrix factorization (nmf). *ASONAM '09*, pages 320–323, July 2009.

[4] M.W. Berry and M. Browne. Email surveillance using non-negative matrix factorization. *Computational & Mathematical Organization Theory*, 11(3):249–264, 2005.

[5] R. Caruana. Multitask learning. *Machine Learning*, 28(1):41–75, 1997.

[6] A.P. Dempster, N.M. Laird, D.B. Rubin, et al. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977.

[7] L. Fei-Fei, R. Fergus, and P. Perona. One-shot learning of object categories. *PAMI*, 28(4):594–611, 2006.

[8] S.A. Golder and B.A. Huberman. Usage patterns of collaborative tagging systems. *Journal of Information Science*, 32(2):198, 2006.

[9] D.R. Hardoon, S. Szedmak, and J. Shawe-Taylor. Canonical correlation analysis: an overview with application to learning methods. *Neural Computation*, 16(12):2639–2664, 2004.

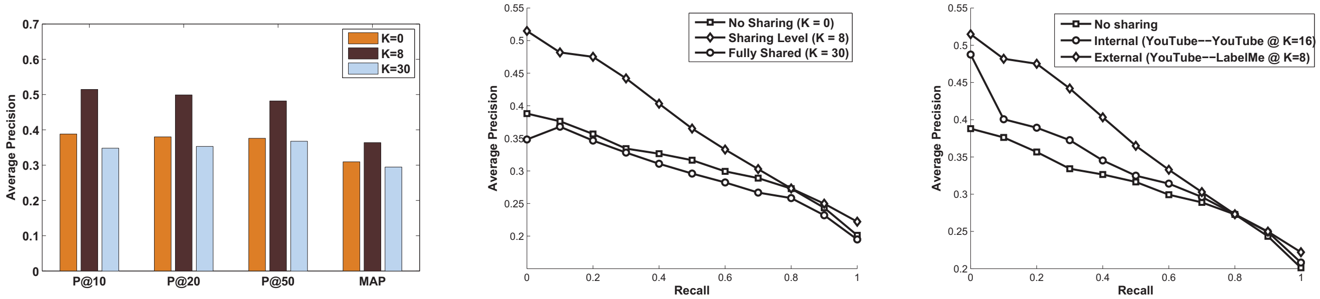
[10] P.O. Hoyer. Non-negative matrix factorization with sparseness constraints. *The Journal of Machine Learning Research*, 5:1457–1469, 2004.

[11] M.S. Kankanhalli and Y. Rui. Application potential of multimedia information retrieval. *Proceedings of the IEEE*, 96(4):712, 2008.

[12] J.R. Kettenring. Canonical analysis of several sets of variables. *Biometrika*, 58(3):433–451, 1971.

[13] D.D. Lee and H.S. Seung. Algorithms for non-negative matrix factorization. In *Advances in Neural Information Processing*, 2000.

[14] X. Li, C. G. M. Snoek, and M. Worring. Learning social tag relevance by neighbor voting. *IEEE Transactions on Multimedia*, in press, 2009.



(a) (b) (c)

Figure 3: YouTube video retrieval results (a) Precision-Scope and MAP plots for $BaselineI$ ($K = 0$), optimally shared subspace ($K = 8$) and $BaselineII$ ($K = 30$) (b) 11-Point interpolated Precision-Recall Curve for $BaselineI$ ($K = 0$), optimally shared subspace ($K = 8$) and $BaselineII$ ($K = 30$) (c) Comparison with case when auxiliary data is chosen from within YouTube domain instead of LabelMe.

[15] X. Li, C.G.M. Snoek, and M. Worring. Annotating images by harnessing worldwide user-tagged photos. *ICASSP. Taipei, Taiwan*, 2009.

[16] Y.R. Lin, H. Sundaram, M. De Choudhury, and A. Kelliher. Temporal patterns in social media streams: Theme discovery and evolution using joint analysis of content and context. In *ICME 2009*, pages 1456–1459, 2009.

[17] K. V. Mardia, J. M. Bibby, and J. T. Kent. *Multivariate analysis*. Academic Press, New York, 1979.

[18] C. Marlow, M. Naaman, D. Boyd, and M. Davis. Ht06, tagging paper, taxonomy, flickr, academic article, toread. *Proceedings of the seventeenth Conference on Hypertext and Hypermedia*, pages 31–40, 2006.

[19] S.J. Pan and Q. Yang. A survey on transfer learning. *Technical Report HKUST-CS08-08, Department of Computer Science and Engineering, HKUST, Hong Kong, China*, 2008.

[20] R. Raina, A. Battle, H. Lee, B. Packer, and A.Y. Ng. Self-taught learning: Transfer learning from unlabeled data. *Proceedings of the 24th International Conference on Machine Learning*, page 766, 2007.

[21] B.C. Russell, A. Torralba, K.P. Murphy, and W.T. Freeman. Labelme: a database and web-based tool for image annotation. *International Journal of Computer Vision*, 77(1):157–173, 2008.

[22] G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5):513–523, 1988.

[23] F. Shahnaz, M.W. Berry, V.P. Pauca, and R.J. Plemmons. Document clustering using nonnegative matrix factorization. *Information Processing and Management*, 42(2):373–386, 2006.

[24] B. Sigurbjörnsson and R. Van Zwol. Flickr tag recommendation based on collective knowledge. *Proceeding of ACM International World Wide Web Conference*, 2008.

[25] C. Wang, F. Jing, L. Zhang, and H.J. Zhang. Scalable search-based image annotation. *Multimedia Systems*, 14(4):205–220, 2008.

[26] X. Wang, C. Pal, and A. McCallum. Generalized component analysis for text with heterogeneous attributes. *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, page 803, 2007.

[27] L. Wu, L. Yang, N. Yu, and X.S. Hua. Learning to tag. *Proceedings of the 18th International Conference on World Wide Web*, pages 361–370, 2009.

[28] Z. Wu, C.W. Cheng, and C. Li. Social and semantics analysis via non-negative matrix factorization. *Proceedings of the 17th International Conference on World Wide Web*, 2008.

APPENDIX

We provide derivations for the optimization problem posed in section 3 which leads to a set of multiplicative updates equations (3)–(7) used in Algorithm 1. Then, we provide a proof of convergence.

A. MULTIPLICATIVE UPDATE EQUATIONS FOR JSNMF

Recall the form of the objective function D to be minimized from section 3:

$$D \triangleq \frac{1}{2}\left\{\left\|\boldsymbol{X} - [\mathbf{W} \mid \boldsymbol{U}]\mathbf{H}\right\|_F^2 + \lambda \left\|\boldsymbol{Y} - [\mathbf{W} \mid \boldsymbol{V}]\mathbf{L}\right\|_F^2\right\}$$

where $\lambda = \|\boldsymbol{X}\|_F^2 / \|\boldsymbol{Y}\|_F^2$. We express D explicitly in terms of the elements of the involved matrices to get:

$$D = \frac{1}{2}\sum_{n=1}^{N_1}\sum_{m=1}^{M}\left[\boldsymbol{X}_{mn} - \sum_{j=1}^{K}\mathbf{W}_{mj}\mathbf{H}_{jn} - \sum_{j=K+1}^{R_1}\boldsymbol{U}_{mj}\mathbf{H}_{jn}\right]^2 + \frac{\lambda}{2}\sum_{n=1}^{N_2}\sum_{m=1}^{M}\left[\boldsymbol{Y}_{mn} - \sum_{j=1}^{K}\mathbf{W}_{mj}\mathbf{L}_{jn} - \sum_{j=K+1}^{R_2}\boldsymbol{V}_{mj}\mathbf{L}_{jn}\right]^2$$

To minimize D, we take the derivative with respect to $\mathbf{W}_{mi}$, $\boldsymbol{U}_{mi}$, $\boldsymbol{V}_{mi}$, $\mathbf{H}_{in}$ and $\mathbf{L}_{in}$. For example, the derivative with respect to $\mathbf{W}_{mi}$ is given by:

$$\nabla_{\mathbf{W}_{mi}} D = \sum_{n=1}^{N_1}(-\mathbf{H}_{in})\left[\boldsymbol{X}_{mn} - \sum_{j=1}^{K}\mathbf{W}_{mj}\mathbf{H}_{jn} - \sum_{j=K+1}^{R_1}\boldsymbol{U}_{mj}\mathbf{H}_{jn}\right] + \lambda\sum_{n=1}^{N_2}(-\mathbf{L}_{in})\left[\boldsymbol{Y}_{mn} - \sum_{j=1}^{K}\mathbf{W}_{mj}\mathbf{L}_{jn} - \sum_{j=K+1}^{R_2}\boldsymbol{V}_{mj}\mathbf{L}_{jn}\right]$$

In general we can observe a decomposition of the gradient into the shared components (those first terms in the bracket), discriminant components (second terms) and the residuals (last two terms).

To minimize the cost function, we follow the optimization procedure similar to Lee and Seung [13] and derive multiplicative update equations using Gradient-Descent method:

$$\mathbf{W}_{mi}^{r+1} \leftarrow \mathbf{W}_{mi}^{r} + \eta_{\mathbf{W}_{mi}} \left(-\nabla_{\mathbf{W}_{mi}} D\right)$$

Substituted with the above derivative, we obtain the general update rule:

$$\begin{aligned} \mathbf{W}_{mi}^{r+1} \quad \leftarrow \quad & \mathbf{W}_{mi}^{r} + \eta_{\mathbf{W}_{mi}} \left(\left[\sum_{n=1}^{N_1} \mathbf{H}_{in} \boldsymbol{X}_{mn} \right] + \left[\lambda \sum_{n=1}^{N_2} \mathbf{L}_{in} \boldsymbol{Y}_{mn} \right] \right) \\ & - \eta_{\mathbf{W}_{mi}} \left[\sum_{n=1}^{N_1} \mathbf{H}_{in} \left(\sum_{j=1}^{K} \mathbf{W}_{mj} \mathbf{H}_{jn} + \sum_{j=K+1}^{R_1} \boldsymbol{U}_{mj} \mathbf{H}_{jn} \right) \right] \\ & - \eta_{\mathbf{W}_{mi}} \lambda \left[\sum_{n=1}^{N_2} \mathbf{L}_{in} \left(\sum_{j=1}^{K} \mathbf{W}_{mj} \mathbf{L}_{jn} + \sum_{j=K+1}^{R_2} \boldsymbol{V}_{mj} \mathbf{L}_{jn} \right) \right] \end{aligned}$$

To achieve a similar effect of multiplicative updating as in [13], we choose the step-size $\eta_{\mathbf{W}_{mi}}$ as

$$\begin{aligned} 1/\eta_{\mathbf{W}_{mi}} \quad = \quad & \frac{\sum_{n=1}^{N_1} \mathbf{H}_{in} \left(\sum_{j=1}^{K} \mathbf{W}_{mj} \mathbf{H}_{jn} + \sum_{j=K+1}^{R_1} \boldsymbol{U}_{mj} \mathbf{H}_{jn} \right)}{\mathbf{W}_{mi}} \\ & + \lambda \frac{\sum_{n=1}^{N_2} \mathbf{L}_{in} \left(\sum_{j=1}^{K} \mathbf{W}_{mj} \mathbf{L}_{jn} + \sum_{j=K+1}^{R_2} \boldsymbol{V}_{mj} \mathbf{L}_{jn} \right)}{\mathbf{W}_{mi}} \end{aligned}$$

Substituting into the previous expression the update equation for $\mathbf{W}_{mi}$ then can be obtained as:

$$\mathbf{W}_{mi}^{r+1} \leftarrow \mathbf{W}_{mi}^{r} \mathbf{S}_{mi}$$

where $\mathbf{S}_{mi}$ is given by (compact form is given in section 3 after Eq. (3))

$$\begin{aligned} 1/\mathbf{S}_{mi} \quad = \quad & \frac{\sum_{n=1}^{N_1} \mathbf{H}_{in} \left(\sum_{j=1}^{K} \mathbf{W}_{mj} \mathbf{H}_{jn} + \sum_{j=K+1}^{R_1} \boldsymbol{U}_{mj} \mathbf{H}_{jn} \right)}{\left[\sum_{n=1}^{N_1} \mathbf{H}_{in} \boldsymbol{X}_{mn} \right] + \left[\lambda \sum_{n=1}^{N_2} \mathbf{L}_{in} \boldsymbol{Y}_{mn} \right]} \\ & + \lambda \frac{\sum_{n=1}^{N_2} \mathbf{L}_{in} \left(\sum_{j=1}^{K} \mathbf{W}_{mj} \mathbf{L}_{jn} + \sum_{j=K+1}^{R_2} \boldsymbol{V}_{mj} \mathbf{L}_{jn} \right)}{\left[\sum_{n=1}^{N_1} \mathbf{H}_{in} \boldsymbol{X}_{mn} \right] + \left[\lambda \sum_{n=1}^{N_2} \mathbf{L}_{in} \boldsymbol{Y}_{mn} \right]} \end{aligned}$$

Similarly, derivatives of D with respect to $\boldsymbol{U}_{mi}$, $\mathbf{H}_{in}$ ($\boldsymbol{V}_{mi}$ and $\mathbf{L}_{in}$ can be written by symmetry) are:

$$\nabla_{\boldsymbol{U}_{mi}} D = \sum_{n=1}^{N_1} \mathbf{H}_{in} \left(\sum_{j=1}^{K} \mathbf{W}_{mj} \mathbf{H}_{jn} + \sum_{j=K+1}^{R_1} \boldsymbol{U}_{mj} \mathbf{H}_{jn} \right) - \sum_{n=1}^{N_1} \mathbf{H}_{in} \mathbf{X}_{mn}$$

$$\nabla_{\mathbf{H}_{in}} D = \sum_{m=1}^{M} \mathbf{W}_{mi} \left(\sum_{j=1}^{K} \mathbf{W}_{mj} \mathbf{H}_{jn} + \sum_{j=K+1}^{R_1} \boldsymbol{U}_{mj} \mathbf{H}_{jn} \right) - \sum_{m=1}^{M} \mathbf{W}_{in} \mathbf{X}_{mn}$$

Similar to the case of $\mathbf{W}_{mi}$, we choose appropriate step-size in each case for $\boldsymbol{U}_{mi}, \boldsymbol{V}_{mi}, \mathbf{H}_{in}$ and $\mathbf{L}_{in}$ to obtain the update equations in (4)–(7). In particular, the following step-sizes for $\boldsymbol{U}_{mi}, \mathbf{H}_{in}$ are chosen (step sizes for $\boldsymbol{V}_{mi}$ and $\mathbf{L}_{in}$ can be written by symmetry):

$$\begin{aligned} \eta_{\boldsymbol{U}_{mi}} \quad &= \quad \frac{\boldsymbol{U}_{mi}}{\sum_{n=1}^{N_1} \mathbf{H}_{in} \left(\sum_{j=1}^{K} \mathbf{W}_{mj} \mathbf{H}_{jn} + \sum_{j=K+1}^{R_1} \boldsymbol{U}_{mj} \mathbf{H}_{jn} \right)} \\ \eta_{\boldsymbol{V}_{mi}} \quad &= \quad \frac{\boldsymbol{V}_{mi}}{\sum_{n=1}^{N_2} \mathbf{L}_{in} \left(\sum_{j=1}^{K} \mathbf{W}_{mj} \mathbf{L}_{jn} + \sum_{j=K+1}^{R_2} \boldsymbol{V}_{mj} \mathbf{L}_{jn} \right)} \end{aligned}$$

B. PROOF OF CONVERGENCE

The proof of convergence makes use of an auxiliary upper bound function similar to the auxiliary lower bound used in the EM-algorithm [6] and extends the proof of convergence of basic NMF given in [13] to Joint Shared NMF (JSNMF) case.

Using the auxiliary function defined in [13], $G(w, w')$ is defined as an upper bound function for $D(w)$ if $G(w, w') \geqslant D(w)$ and equality is satisfied iff $w = w'$. Note that minimizing the upper bound function G at every update of w leads to non-increasing function D on every update. Hence if $w^{t+1} = \operatorname*{argmin}_{w} G(w^{t+1}, w^{t})$, then

$$D\left(w^{t+1}\right) \leq G\left(w^{t+1}, w^{t}\right) \leq G\left(w^{t}, w^{t}\right) = D\left(w^{t}\right)$$

Also note that when $D\left(w^{t+1}\right) = D\left(w^{t}\right)$, it implies that w^{t} is a local minimum of $G\left(w, w^{t}\right)$ and if derivatives of D exist and are continuous in a small neighborhood $\left\|w^{t} - \delta w\right\| < \epsilon_0$, this also implies that $\nabla_{w} D\left(w^{t}\right) = 0$. Denote $\mathbf{1}_{a,b}$ to be the identity function, i.e., return 1 if $a = b$ and 0 otherwise, we shall prove the following lemma extended from [13] to our JSNMF case:

Lemma. *If $K(w_i)$ is the diagonal matrix with its $(a,b)^{th}$element given by*

$$K_{ab}(w_i) = \mathbf{1}_{a,b} \frac{\left(\mathbf{H}_w \mathbf{H}_w^{\mathsf{T}} w_i + \mathbf{H}_w \mathbf{H}_u^{\mathsf{T}} u_i\right)_a + \lambda \left(\mathbf{L}_w \mathbf{L}_w^{\mathsf{T}} w_i + \mathbf{L}_w \mathbf{L}_v^{\mathsf{T}} v_i\right)_a}{\left(w_i\right)_a}$$

then

$$\begin{aligned} G\left(w_i, w_i^{t}\right) \quad &= \quad D\left(w_i^{t}\right) + \left(w_i - w_i^{t}\right)^{\mathsf{T}} \nabla_{w_i} D\left(w_i^{t}\right) \\ & \quad + \frac{1}{2}\left(w_i - w_i^{t}\right)^{\mathsf{T}} K\left(w_i^{t}\right)\left(w_i - w_i^{t}\right) \end{aligned}$$

is an auxiliary function for

$$\begin{aligned} D\left(w_i\right) \quad = \quad & \frac{1}{2} \sum_{i} \{ \left(x_i^{\mathsf{T}} - \left[w_i^{\mathsf{T}} \mid u_i^{\mathsf{T}}\right] \mathbf{H}\right) \left(x_i^{\mathsf{T}} - \left[w_i^{\mathsf{T}} \mid u_i^{\mathsf{T}}\right] \mathbf{H}\right)^{\mathsf{T}} \\ & + \lambda \left(y_i^{\mathsf{T}} - \left[w_i^{\mathsf{T}} \mid v_i^{\mathsf{T}}\right] \mathbf{L}\right) \left(y_i^{\mathsf{T}} - \left[w_i^{\mathsf{T}} \mid v_i^{\mathsf{T}}\right]\right)^{\mathsf{T}} \} \end{aligned}$$

where x_i, w_i, u_i and v_i are the row vectors of matrices $\boldsymbol{X}$, $\mathbf{W}$, $\boldsymbol{U}$ and $\boldsymbol{V}$ respectively and $\mathbf{H} \triangleq \left[\mathbf{H}_w^{\mathsf{T}} \mid \mathbf{H}_u^{\mathsf{T}}\right]^{\mathsf{T}}$, $\mathbf{L} \triangleq \left[\mathbf{L}_w^{\mathsf{T}} \mid \mathbf{L}_v^{\mathsf{T}}\right]^{\mathsf{T}}$.

PROOF. **T**he first and second derivative of $F\left(w_i\right)$ are given by

$$\begin{aligned} \nabla_{w_i} D\left(w_i\right) &= -\mathbf{H}_w x_i + \mathbf{H}_w \mathbf{H}_w^{\mathsf{T}} w_i + \mathbf{H}_w \mathbf{H}_u^{\mathsf{T}} u_i \\ & \quad + \lambda \left(-\mathbf{L}_w y_i + \mathbf{L}_w \mathbf{L}_w^{\mathsf{T}} w_i + \mathbf{L}_w \mathbf{L}_v^{\mathsf{T}} v_i\right) \\ \nabla_{w_i}^2 D\left(w_i\right) &= \mathbf{H}_w \mathbf{H}_w^{\mathsf{T}} + \lambda \mathbf{L}_w \mathbf{L}_w^{\mathsf{T}} \end{aligned}$$

Comparing $D\left(w_i\right)$with $G\left(w_i, w_i^{t}\right)$, we see that all we need to prove is the following

$$\left(w_i - w_i^{t}\right)^{\mathsf{T}} \left(K\left(w\right) - \mathbf{H}_w \mathbf{H}_w^{\mathsf{T}} - \lambda \mathbf{L}_w \mathbf{L}_w^{\mathsf{T}}\right)\left(w_i - w_i^{t}\right) \geq 0$$

To prove the positive semi-definiteness, consider the matrix with elements

$$M_{ab}\left(w^{t}\right) = \left(w_i^{t}\right)_a \left(K\left(w^{t}\right) - \mathbf{H}_w \mathbf{H}_w^{\mathsf{T}} - \lambda \mathbf{L}_w \mathbf{L}_w^{\mathsf{T}}\right)_{ab} \left(w_i^{t}\right)_b$$

which is just a rescaling of the elements of matrix $K\left(w\right) - \mathbf{H}_w \mathbf{H}_w^{\mathsf{T}} - \lambda \mathbf{L}_w \mathbf{L}_w^{\mathsf{T}}$. Then $K\left(w\right) - \mathbf{H}_w \mathbf{H}_w^{\mathsf{T}} - \lambda \mathbf{L}_w \mathbf{L}_w^{\mathsf{T}}$ is positive semi-definite if and only if M is. Equivalently we need to prove $\nu^{\mathsf{T}} M \nu \geq 0, \forall \nu$.

This is indeed the case. By explicitly expressing $\nu^\intercal M\nu$, we can show that it is a sum of nonnegative terms. To avoid lengthy derivation, we only state the main results here:

$$\begin{aligned}
\nu^\intercal M\nu &= \sum_{a,b} \nu_a \left(w_i^t\right)_a \left(K\left(w^t\right) - \mathbf{H}_w\mathbf{H}_w^\intercal - \lambda\mathbf{L}_w\mathbf{L}_w^\intercal\right)_{ab} \left(w_i^t\right)_b \nu_b \\
&= \sum_{a,b} \left(w_i^t\right)_a \left(\mathbf{H}_w\mathbf{H}_w^\intercal\right)_{ab} \left(w_i^t\right)_b \frac{\left(\nu_a - \nu_b\right)^2}{2} \\
&+ \lambda \sum_{a,b} \left(w_i^t\right)_a \left(\mathbf{L}_w\mathbf{L}_w^\intercal\right)_{ab} \left(w_i^t\right)_b \frac{\left(\nu_a - \nu_b\right)^2}{2} \\
&+ \sum_{a,b_2} \nu_a^2 \left(\mathbf{H}_w\mathbf{H}_u^\intercal\right)_{ab_2} \left(w_i^t\right)_a \left(w_i^t\right)_{b_2} \\
&+ \lambda \sum_{a,b_4} \nu_a^2 \left(\mathbf{L}_w\mathbf{L}_v^\intercal\right)_{ab_4} \left(w_i^t\right)_a \left(w_i^t\right)_{b_4} \\
&\geq 0
\end{aligned}$$

Back to our main proof of convergence, the gradient of the auxiliary function G is given by

$$\nabla_{w_i} G\left(w_i, w_i^t\right) = \nabla_{w_i} D\left(w_i^t\right) + K\left(w_i^t\right)\left(w_i - w_i^t\right)$$

To obtain the local minimum of $G\left(w_i, w_i^t\right)$, we equate $\nabla_{w_i} G\left(w_i, w_i^t\right)$ to zero and get the following update equation :

$$w_i^{t+1} = w_i^t - K^{-1}\left(w_i^t\right)\nabla_{w_i} D\left(w_i^t\right)$$

Comparing the above update equation to the Gradient- Descent update equation, we get the following step-size, $\eta_{w_i^t}$

$$\eta_{w_i^t} = K^{-1}\left(w_i^t\right)$$

For brevity, hereafter we shall drop the update iteration superscript $t, t+1$ and use $\leftarrow$notation to denote the update, i.e.,

$$w_i \leftarrow w_i - K^{-1}\left(w_i\right)\nabla_{w_i} D\left(w_i\right)$$

After substituting for $\nabla_{w_i} D\left(w_i\right)$, we get the following update for w_i

$$\begin{aligned}
\left(w_i\right)_a \quad \leftarrow \quad & \left(w_i\right)_a - [\left(-x_i^\intercal\mathbf{H}_w^\intercal + w_i^\intercal\mathbf{H}_w\mathbf{H}_w^\intercal + u_i^\intercal\mathbf{H}_u\mathbf{H}_w^\intercal\right)_a \\
& +\lambda\left(-y_i^\intercal\mathbf{L}_w^\intercal + w_i^\intercal\mathbf{L}_w\mathbf{L}_w^\intercal + v_i^\intercal\mathbf{L}_v\mathbf{L}_w^\intercal\right)_a] \\
& \frac{\left(w_i\right)_a}{\left(\mathbf{H}_w\mathbf{H}_w^\intercal w_i + \mathbf{H}_w\mathbf{H}_u^\intercal u_i\right)_a + \lambda\left(\mathbf{L}_w\mathbf{L}_w^\intercal w_i + \mathbf{L}_w\mathbf{L}_v^\intercal v_i\right)_a}
\end{aligned}$$

Now, the above expression can be written in terms of matrix notation as the following :

$$(\mathbf{W})_{ab} \leftarrow (\mathbf{W})_{ab} \frac{\left(\boldsymbol{X}\mathbf{H}_w^\intercal + \lambda\boldsymbol{Y}\mathbf{L}_w^\intercal\right)_{ab}}{\left(\mathbf{W}\mathbf{H}_w\mathbf{H}_w^\intercal + \boldsymbol{U}\mathbf{H}_u\mathbf{H}_w^\intercal\right)_{ab} + \lambda\left(\mathbf{W}\mathbf{L}_w\mathbf{L}_w^\intercal + \boldsymbol{V}\mathbf{L}_v\mathbf{L}_w^\intercal\right)_{ab}}$$

Similar to the case of w_i, if we choose

$$K_{ab}\left(u_i\right) = \mathbf{1}_{a,b}\frac{\left(\mathbf{H}_u\mathbf{H}_w^\intercal w_i + \mathbf{H}_u\mathbf{H}_u^\intercal u_i\right)_a}{\left(u_i\right)_a}$$

Then we can prove that $G\left(u_i, u_i^t\right)$ is an auxiliary function for $D\left(u_i\right)$. The first and second derivatives are given by

$$\begin{aligned}
&\nabla_{u_i} D\left(u_i\right) = -\mathbf{H}_u x_i + \mathbf{H}_u\mathbf{H}_w^\intercal w_i + \mathbf{H}_u\mathbf{H}_u^\intercal u_i \\
&\nabla_{u_i}^2 D\left(u\right) = \mathbf{H}_u\mathbf{H}_u^\intercal
\end{aligned}$$

Comparing $D\left(u_i\right)$with $G\left(u_i, u_i^t\right)$, we see that all we need to prove is the following

$$\left(u_i - u_i^t\right)^\intercal\left(K\left(u\right) - \mathbf{H}_u\mathbf{H}_u^\intercal\right)\left(u_i - u_i^t\right) \geq 0$$

The proof for this case is exactly similar to that of the case of w_i. Doing that, we get the gradient-descent step size as follows

$$\eta_{u_i^t} = K^{-1}\left(u_i^t\right) = \frac{\left(u_i^t\right)_a}{\left(\mathbf{H}_u\mathbf{H}_w^\intercal w_i + \mathbf{H}_u\mathbf{H}_u^\intercal u_i^t\right)_a}$$

and the updates for $\boldsymbol{U}$ (update expression for $\boldsymbol{V}$ can be written by symmetry) as below

$$(\boldsymbol{U})_{ab} \leftarrow (\boldsymbol{U})_{ab}\frac{\left(\boldsymbol{X}\mathbf{H}_u^\intercal\right)_{ab}}{\left(\mathbf{W}\mathbf{H}_w\mathbf{H}_u^\intercal + \boldsymbol{U}\mathbf{H}_u\mathbf{H}_u^\intercal\right)_{ab}}$$

The update expression for $\boldsymbol{V}$ can be written by symmetry between $\boldsymbol{U}$ and $\boldsymbol{V}$ and those for $\mathbf{H}$ and $\mathbf{L}$ are similar to basic NMF updates.