Probabilistic Matrix Factorization With Semantic And Visual Neighborhoods For Image Tag Completion

Dimitrios Rafailidis
Department of Informatics
Aristotle University, Thessaloniki, Greece
draf@csd.auth.gr

ABSTRACT

We present an image tag completion method, namely PMF-SVN, where the key idea is to exploit images' Semantically and Visually similar Neighborhoods (SVNs) in the learning process of a Probabilistic Matrix Factorization (PMF) framework. We propose a two-step SVN formation algorithm that can generate an image set with the images being both visually and semantically similar. Furthermore, we introduce an efficient way to incorporate the formed SVNs into the learning process of PMF, under the constraint that the latent features of each image are averaged by the features of the images that belong to its SVN. In our experiments with benchmark datasets, we show that the proposed PMF-SVN method outperforms competitive baselines, in terms of completion accuracy, by efficiently capturing the semantical and visual associations between images and tags in SVNs.

Categories and Subject Descriptors

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

General Terms

Algorithms, Measurement, Experimentation

Keywords

Tag Completion, Image Annotation, Matrix Factorization

1. INTRODUCTION

With the advent of social media platforms in the past decade, image tagging has gained a lot of attraction by researchers. Tags are provided by users in the form of free text and they are usually imprecise, containing noise to efficiently describe the visual content of images [2, 7]. Meanwhile, users often avoid assigning tags to images, making image tags incomplete. To deal with noisy and incomplete tags, several completion methods [2, 3, 7] have been recently proposed, to capture relevant tag-image associations and consequently to add the missing tags.

ICMR'15 June 23-26, 2015, Shanghai, China.

http://dx.doi.org/XXX

Although tag completion methods have several applications such as image retrieval [7], event detection [1], etc., several real world problems limit their performance. The different tag choices introduce several problems; for example, a tag with multiple meanings (polysemy), different tags with similar meanings (synonymy), as well as using general versus specialized tags to refer to the same concept, just to name a few. Matrix factorization techniques, such as probabilistic matrix factorization (PMF), Bayesian PMF, non-negative matrix factorization and trace norm regularized matrix factorization, are suitable for facing the aforementioned problems, since such techniques can reveal the latent tag-image associations and complete the missing tags, accordingly. However, in many cases their performance is negatively affected by the very sparse annotations of images, since in this case there are only few annotations on which to base the generation of the latent tag-image associations and consequently the completion of relevant tags is limited. To solve the problem of images' sparse annotations, several tag completion methods [2, 3, 7] exploit the visual features to construct relevant image-tag associations by forming representations of the missing tags of an image using the tags of its visually similar images. However, this may also limit the performance of tag completion, since images that are visually similar are not necessarily semantically similar. What is therefore required is an efficient way to capture the associations between visually and semantically similar images.

Our contribution is summarized as follows, **(C1:)** we propose an efficient algorithm to generate for each image a set of semantically and visually similar images based on the textual information of tags and the visual features of images, forming thus a Semantically and Visually similar Neighborhood (SVN) for each image; **(C2:)** we introduce an efficient way to incorporate the formed SVNs into the learning process of PMF to solve the image tag completion problem.

2. RELATED WORK

Image auto-annotation methods [5] are related to the image tag completion problem; however, these methods differ, since the majority of existing auto-annotation methods assume that images in training set are completely annotated with appropriate tags [2, 3], whereas tag completion methods add missing relevant tags, given a dataset made up of partially annotated images. Also, several tag refinement methods have been proposed, such as the work of [8], focused more on removing noisy tags but less on completion.

More relevant to our work, Wu et al. [7] proposed to address the tag completion problem by searching for the optimal tagging matrix consistent with both tags and visual

similarities. Recently, Lin et al. [3] introduced an image tag completion method via image-specific and tag-specific Linear Sparse Reconstruction (LSR), which reconstructs each image and each tag with the remaining ones using a sparse coding technique. Then, both image-specific and tag-specific reconstructions are merged, in order to complete missing relevant tags. Feng et al. [2] proposed a tag completion algorithm (TCMR) which aims to simultaneously enrich the missing tags and remove noisy ones, by exploiting the statistical dependence between image features and tags via a graph Laplacian. The impact of incomplete and noisy tags was reduced by assigning high weights to tags that are consistent with image features, and low weights otherwise. Compared to state-of-the-art, the proposed method captures the inter-correlations between images on the condition that images are both visually and semantically similar in order to form SVNs, handling thus noisy and incomplete image tags in a probabilistic matrix factorization framework.

3. PROPOSED PMF-SVN

3.1 Problem Definition & PMF

Problem Definition: Given N tags and M images, initially we calculate a matrix $R = [R_{t,i}] \in \mathbb{R}^{N \times M}$, where each element $R_{t,i}$=1 denotes the degree of relevance (tag assignment) between tag t and image i, and $R_{t,i}$=0 otherwise. Matrix R is usually very sparse, since most of its elements are expected to be missing in the tag completion problem. Also, we compute a similarity matrix $S = [S_{i,j}] \in \mathbb{R}^{M \times M}$ based on a visual similarity function $f_V(\cdot)$, with $S_{i,j}$=$f_V(i,j)$ (Section 3.2). The image tag completion problem is defined as follows: given matrix R with the existing tag assignments between N tags and M images, as well as the visual similarities of the M images in S, the goal is to complete the relevant missing tag-image associations in R.

Probabilistic Matrix Factorization (PMF): matrix factorization methods (a.k.a low rank D approximations), such as the PMF model [6], construct a latent-feature D-dimensional space in which they represent each tag and image. Let $U \in \mathbb{R}^{D \times N}$ be a matrix whose t-th column vector, denoted as U_t, represents the t-th tag in the D-dimensional space. Similarly, let $V \in \mathbb{R}^{D \times M}$ matrix whose i-th column vector, denoted as V_i, represents the i-th image in the same D-dimensional space. Using the initial matrix $R \in \mathbb{R}^{N \times M}$ as training data, matrix factorization techniques learn, i.e. compute the elements of matrices U and V, so that they can approximate matrix R with matrix $\widehat{R}$, such that $R \approx \widehat{R} = U^T V$. The process of learning matrices U and V can be expressed in the probabilistic framework of PMF, where the likelihood of observing a specific tag-image association in R can be expressed as:

$$p\big(R|U,V;\sigma_R^2\big) = \prod_{t=1}^{N}\prod_{i=1}^{M}\big[\mathfrak{N}\big(R_{t,i}|U_t^T V_i;\sigma_R^2\big)\big]^{\mathbf{1}_R(t,i)} \quad (1)$$

where $\mathfrak{N}(\mu,\sigma^2)$ denotes the normal distribution with mean μ and variance σ^2; and $\mathbf{1}_R(t,i)$ denotes the indicator function with 1 if $R_{t,i}$ is known, i.e. not null, or 0 otherwise. Eq. (1) makes the premise that each known relation, represented with the element $R_{t,i}$ is an independent and identically distributed *(iid)* random variable that follows a normal distribution whose mean values is equal to $\widehat{R}_{t,i} = U_t^T V_i$ and whose variance σ^2 is treatment as hyperparameter. Based on the Bayes theorem and the likelihood function of Eq. (1), we can obtain the posterior probability of U and V:

$$p\big(U,V|R;\sigma_R^2,\sigma_U^2,\sigma_V^2\big) \propto \prod_{t=1}^{N}\prod_{i=1}^{M}\big[\mathfrak{N}\big(R_{t,i}V_i;\sigma_R^2\big)\big]^{\mathbf{1}_R(t,i)} \times \ldots \times \prod_{t=1}^{N}\big[\mathfrak{N}\big(U_t|0;\sigma_U^2\big)\big] \times \prod_{i=1}^{M}\big[\mathfrak{N}\big(V_i|0;\sigma_V^2\big)\big] \quad (2)$$

where Eq. (2) makes the premise that the features U_t of each tag t , as well as the features V_i of each image i in the D-dimensional space, are also *iid* random variables following normal distribution with zero mean and variances σ_U^2 and σ_V^2, respectively. To compute the low rank D approximation $\widehat{R}$, we have to calculate U and V, so as to maximize the probability of observing the given tag-image associations in R. We pose the image tag completion problem as a problem of maximizing the posterior probability in Eq. (2). The proposed PMF-SVN method consists of the following two steps, (i) for the M images, M respective SVNs are constructed based on the SVN formation algorithm of Section 3.2; (ii) the M constructed SVNs are incorporated into the learning process of PMF, by transforming the maximization problem of Eq. (2) into a minimization problem, which is furthermore solved using the optimization algorithm of gradient descent.

3.2 SVN Formation Algorithm

The inputs of the SVN formation algorithm are: (i) an image q; (ii) the tag set $\mathcal{T}_q$, i.e. the set of tags assigned to q based on R; (iii) a visual threshold ϵ. The algorithm returns the semantically and visually similar neighborhood (SVN) of image q as set $\mathcal{N}^+(q)$. In the SVN formation algorithm the following functions are defined:

(i) *Visual similarity function $f_V(\cdot)$:* a bag-of-words model based on densely sampled SIFT descriptors is used to represent the visual content, and $f_V(\cdot)$ is generated by normalizing the $L2$-norm[1] of image features in [0 1], transformed to similarity with $S_{i,j} = f_V(i,j)$.

(ii) *Aggregated visual similarity function $f_A(\cdot)$:* is computed between an image set $\mathcal{I}$ and an image x as follows:

$$f_A(\mathcal{I},x) = \sum_{i \in \mathcal{I}} f_V(i,x) \quad (3)$$

(iii) *Tag similarity function $f_T(\cdot)$:* is calculated according to the Dice similarity [4], ranging in [0 1]:

$$f_T(t_x,t_k) = \frac{2 \cdot |\mathcal{I}_{t_x} \cap \mathcal{I}_{t_k}|}{|\mathcal{I}_{t_x}| + |\mathcal{I}_{t_k}|} \quad (4)$$

where $\mathcal{I}_{t_x}$ and $\mathcal{I}_{t_k}$ denote the sets of images that have been annotated with tags t_x and t_k, respectively.

(iv) *Sum-of-Squared Error function $SSE(\cdot)$* : is calculated between a tag t_x and a tag set $\mathcal{T}$, also in the range of [0 1]:

$$SSE(t_x,\mathcal{T}) = \frac{1}{|\mathcal{T}|} \cdot \sum_{t_k \in \mathcal{T}} \Big(1 - f_T(t_x,t_k)\Big)^2 \quad (5)$$

Initialization step: In line 4 of Algorithm 1, image q is posed as query to generate the set $\mathcal{N}(q)$ with the visual neighbors r based on the visual similarity function $f_V(\cdot)$, with $f_V(q,r) \geq \epsilon$, $\forall\ r \in \mathcal{N}(q)$. In line 5, $\forall$ tag $t_x \in \mathcal{T}_q$ the SSE function $SSE(t_x,\mathcal{T}_q)$ is calculated to identify the tag

[1]Except SIFT descriptors and $L2$-norm several alternatives can be easily used in our method.

t_m of image query q that generates the minimum SSE[2]:

$$t_m = \arg\min_{t_x \in \mathcal{T}_q} SSE(t_x, \mathcal{T}_q) \quad (6)$$

In line 6, we set a a threshold SSE_{thres} equal to the minimum SSE of the identified tag $t_m \in \mathcal{T}_q$. Next, in lines 7-13 each image $r \in \mathcal{N}(q)$ is inserted to the set $\mathcal{N}^+(q)$, if the following condition is satisfied:

$$SSE(t_m, \mathcal{T}_{r_j}) \leq SSE_{thres} \quad (7)$$

where $\mathcal{T}_{r_j}$ is the tag set of the visual neighbor r_j. Initialization finishes with the update of SSE_{thres}, in line 14.

Iterative step: In line 18, the previously identified set $N^+(q)$ is considered as a new query set $\mathcal{Q} \leftarrow \mathcal{N}^+(q)$. According to the aggregated visual similarity function $f_A(\mathcal{Q}, r)$ of Eq.(3), a new set $\mathcal{N}(Q)$ of visual neighbors is generated in line 19. Next, similar to the initialization step, the visual neighbors $r \in \mathcal{N}(Q)$ are examined on the condition of the SSE_{thres} in line 22, where each r is inserted into the set $\mathcal{N}^+(q)$, accordingly. In line 27, threshold SSE_{thres} is updated, which is either preserved or becomes stricter for the next iteration, considering that the visual similarity to the initial query q declines over the iterations (by generating a new query set Q in each iteration) and thus, an equal or stricter SSE_{thres} is required when a new iteration starts. The iterative step terminates if the set $\mathcal{N}^+(q)$ is preserved and the SVN of query q is returned as the final set $\mathcal{N}^+(q)$.

ALGORITHM 1: SVN formation algorithm

Input: (1) image q; (2) tag set $\mathcal{T}_q$; (3) visual threshold ϵ
Output: SVN set $\mathcal{N}^+(q)$.

1
2 **Initialization step**
3 $\mathcal{N}^+(q) \leftarrow \varnothing$;
4 Calculate $\mathcal{N}(q) = \{r_1, r_2, \ldots, r_{|\mathcal{N}(q)|}\}$, with $f_V(q, r_j) \geq \epsilon$;
5 Identify tag t_m with the minimum SSE based on (6);
6 Set SSE_{thres}=$SSE(t_m, \mathcal{T}_q)$;
7 **for** $(j = 1 : |\mathcal{N}(q)|)$;
8 **do**
9 **if** $(SSE(t_m, \mathcal{T}_{r_j}) \leq SSE_{thres})$;
10 **then**
11 Update $\mathcal{N}^+(q) \leftarrow \{\mathcal{N}^+(q), r_j\}$;
12 **end**
13 **end**
14 Update $SSE_{thres} = \min_{\forall r_j \in \mathcal{N}^+(q)} \Big(SSE(t_m, \mathcal{T}_{r_j})\Big)$;
15
16 **Iterative step**
17 **repeat**
18 Set $\mathcal{Q} \leftarrow \mathcal{N}^+(q)$;
19 Calculate $\mathcal{N}(\mathcal{Q})$ based on $f_A(\mathcal{Q}, r) \geq \epsilon$ and (3);
20 **for** $(j = 1 : |\mathcal{N}(\mathcal{Q})|)$;
21 **do**
22 **if** $(SSE(t_m, \mathcal{T}_{r_j}) \leq SSE_{thres})$;
23 **then**
24 Update $\mathcal{N}^+(q) \leftarrow \{\mathcal{N}^+(q), r_j\}$;
25 **end**
26 **end**
27 Update $SSE_{thres} = \min_{\forall r_j \in \mathcal{N}^+(q)} \Big(SSE(t_m, \mathcal{T}_{r_j})\Big)$;
28 **until** $(|\mathcal{N}^+(Q)| = |\mathcal{N}^+(q)|$;
29 **return** $\mathcal{N}^+(q)$;

3.3 Learning PMF With SVNs

Incorporation Of SVNs Into PMF: Given the similarity matrix S with the visual similarities, and the M formed SVNs, let V_i and V_j denote the feature vectors of the i-th and j-th image in the D-dimensional latent-feature space. If image j is in the i-th SVN $\mathcal{N}^+(i)$ of image i, then the learning process of PMF should compute V_i by taken into account V_j. This means that the learning process should consider the semantical and visual information in $\mathcal{N}^+(i)$. This is achieved by firstly expressing the dependency of matrix V on S as:

$$p\big(V|T; \sigma_V^2, \sigma_S^2\big) \propto p\big(V|T\sigma_S^2;\big) \times p\big(V|\sigma_V^2\big) \quad (8)$$

where the first factor $p\big(V|T\sigma_S^2;\big)$ expresses the dependence of V on S and the second factor $p\big(V|\sigma_V^2\big)$ is the prior probability of V. Given the i-th SVN $\mathcal{N}^+(i)$ of image i, we have:

$$p(V|T; \sigma_S^2) = \prod_{i=1}^{M} \Big[\mathfrak{N}\Big(V_i| \sum_{j \in \mathcal{N}^+(i)} S_{i,j}; \sigma_S^2\Big)\Big] \quad (9)$$

Eq. (9) considers that the D-dimensional feature vector V_i follows normal distribution with mean equal to the average of the features of the images that belong to its SVN $\mathcal{N}^+(i)$. Based on Eq. (9) we reformulate the posterior probability of U and V in Eq. (2):

$$p(U, V|R; \sigma_R^2, \sigma_U^2, \sigma_V^2, \sigma_S^2) \propto \prod_{t=1}^{N} \prod_{i=1}^{M} \Big[\mathfrak{N}(R_{t,i}|U_t^T V_i; \sigma_R^2)\Big]^{\mathbf{1}_R(t,i)} \times \quad (10)$$

$$\ldots \times \prod_{t=1}^{N} \Big[\mathfrak{N}(U_t|0; \sigma_U^2)\Big] \times \prod_{i=1}^{M} \Big[\mathfrak{N}(V_i| \sum_{j \in \mathcal{N}^+(i)} S_{i,j} V_j; \sigma_S^2)\Big] \times$$

$$\ldots \times \prod_{i=1}^{M} \Big[\mathfrak{N}(V_i|0; \sigma_V^2)\Big]$$

Objective Function: Eq. (10) provides the basis for learning U and V by exploiting the semantically and visually information of the SVNs. The learning process is performed by calculating those U and V variables that maximize the posterior probability of Eq. (10). Since the natural logarithm function $\ln(p(U, V|R; \sigma_R^2, \sigma_U^2), \sigma_V^2, \sigma_S^2)$ is monotically increasing, we proceed by minimizing its arithmetic-negation function $L(U, V) = -\ln(p(U, V|R; \sigma_R^2, \sigma_U^2), \sigma_V^2, \sigma_S^2)$. Thus, the maximization problem of Eq. (10) is transformed to the following minimization problem:

$$L(U, V) = \frac{1}{2} \sum_{t=1}^{N} \sum_{i=1}^{M} \mathbf{1}_R(t, i)(R_{t,i} - U_t^T V_i)^2 + \frac{\lambda_U}{2} \sum_{t=1}^{N} U_t^T U_t + \quad (11)$$

$$\frac{\lambda_V}{2} \sum_{i=1}^{M} V_i^T V_i + \frac{\lambda_S}{2} \sum_{i=1}^{M} \Big((V_i - \sum_{j \in \mathcal{N}^+(i)} S_{i,j} V_j)^T (V_i - \sum_{j \in \mathcal{N}^+(i)} S_{i,j} V_j)\Big)$$

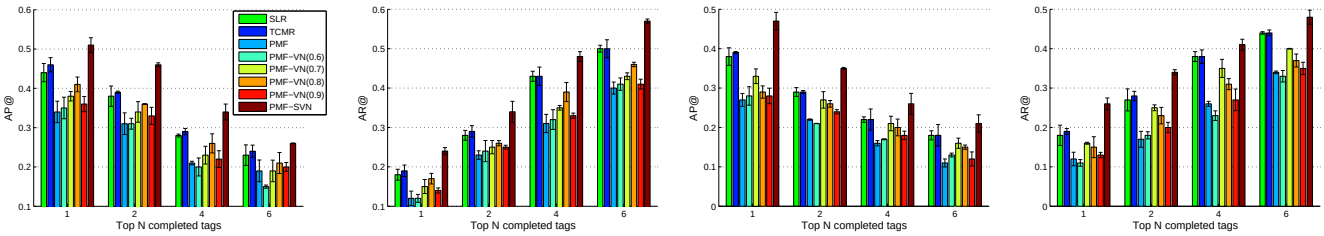
where $\lambda_U = \sigma_R^2/\sigma_U^2$, $\lambda_V = \sigma_R^2/\sigma_V^2$ and $\lambda_S = \sigma_R^2/\sigma_S^2$ are the regularization parameters to avoid model overfitting. To minimize the objective function $L(U, V)$ in Eq. (11), which is a convex function, we use the gradient descent on $\partial L/\partial U_t$ and $\partial L/\partial V_i$ for each pair U_t and V_i, and iteratively update their values. Given a learning rate η, in each iteration (a.k.a. epoch), updating is performed as follows:

$$U_t \leftarrow U_t - \eta \frac{\partial L}{\partial U_t} \ , \ V_i \leftarrow V_i - \eta \frac{\partial L}{\partial V_i} \quad (12)$$

4. EXPERIMENTS

Datasets: In our experiments, we used the benchmark datasets (i) IAPR TC12 from ImageCLEF and (ii) ESP Game, with their features publicly available at [9]. Each dataset was split into training and test set. Following the

[2]Based on Eqs. (4) and (5), t_m with the minimum SSE is the tag with the highest probability of being assigned to more images than the rest of tags in $\mathcal{T}_q$, and thus t_m has the highest probability of being in a tag set $\mathcal{T}_{r_j}$ of a visual neighbor r_j. The mathematical proof is left for future work.



(a) IAPR TC12 (@AP) IAPR TC12 (@AR) (c) ESP game (@AP) (d) ESP game (@AR)

Figure 1: Comparison of PMF-SVN against LSR, TCMR, PMF and PMF-VN(ϵ).

evaluation protocol of [2], each image in the training set has 4 tags. The rest tags of each image were deleted from the training set and used as ground-truth tags in the test set. According to [2, 3], we evaluated the tag completion accuracy by Average Precision ($AP@N$), Average Recall ($AR@N$) and Coverage ($C@N$). $AP@N$ measures the average percentage of the top N correctly completed tags; $AR@N$ the percentage of correct tags that are completed by a method out of all ground-truth tags; and coverage ($C@N$) the percentage of images with at least one correctly completed tag. We ran 20 trials and mean and standard deviation of the evaluation metrics are reported.

Methods: Regarding the proposed PMF-SVN, the learning rate η in Eq. (11) was varied in $[10^{-3}10^{3}]$ with a step of 10, where the concluded η were 10^{-1} and 10^{-2} for IAPR TC12 and ESP game, respectively, denoting that dataset ESP game required a more conservative learning strategy than IAPR TC12. According to Algorithm 1 the proposed PMF-SVN generated SVNs with 7.23 and 11.46 images on average for IAPR TC12 and ESP game. LSR [3] and TCMR [2] were considered as current baselines (Section 2). The parameter values in the baselines were determined by cross-validation, using the publicly available implementations of LSR and TCMR at [10] and [11], respectively. Also, we compared the proposed PMF-SVN against the baseline PMF model [6] which does not exploit any additional information in the learning process by trying to maximize the posterior probability in Eq. (2). Finally, to evaluate the SVN formation Algorithm 1, we also considered as baseline the PMF-VN(ϵ) method, which in contrast to PMF-SVN, uses solely the Visual Neighborhood (VN) $\mathcal{N}(i)$ for each image i, with $f_V(i,r) \geq \epsilon \ \forall r \in \mathcal{N}(i)$, by replacing $\mathcal{N}^+(i)$ with $\mathcal{N}(i)$ in the learning process of Section 3.3.

Results: In Figure 1, the experimental results are presented. In the baseline PMF-VN(ϵ) method, we varied the visual threshold in [0.6 0.7 0.8 0.9]. For PMF-VN(ϵ) the optimal values of the visual threshold ϵ are 0.8, i.e. PMF-VN(0.8) and 0.7, i.e. PMF-VN(0.7) for IAPR TC12 and ESP Game, respectively. The highest visual threshold ϵ=0.9, i.e. PMF-VN(0.9), does not improve the completion accuracy of PMF by limiting the size of the Visual Neighborhood (VN) and consequently having limited impact on PMF; meanwhile the lowest visual threshold ϵ=0.6, i.e. PMF-VN(0.6), generates large visual neighborhoods with less visually similar images, introducing thus noise in PMF. Hence, by solely considering visual neighbors in PMF, there is a glass ceiling in terms of completion accuracy, since baseline PMF-VN(ϵ) does not reach the completion accuracy of LSR and TCMR for any ϵ variation. For the proposed PMF-SVN we consider the same optimal ϵ thresholds, i.e. 0.8 and 0.7 for TAPR TC12 and ESP game. PMF-SVN outperforms PMF and PMF-VN(ϵ), since PMF-SVN based on Algorithm 1 forms SVNs with images that are both semantically and visually similar. Moreover, the proposed PMF-SVN method completes tags more accurate than TCMR and LSR. As also presented in [2], TCMR and LSR have comparable performance in terms of $AP@N$ and $AR@N$ on IAPR TC12 and EPS game; however TCMR achieves higher coverage $C@N$ than LSR. Hence, except the experiments of Fig. 1, we measured the coverage $C@N$, where we observed that PMF-SVN achieves also higher coverage than TCMR; for instance, PMF-SVN achieves $C@2$=[0.7 0.57] in IAPR TC12 and ESP game, whereas TCMR is limited to $C@2$=[0.61 0.51], indicating that PMF-SVN can complete relevant tags for more images than TCMR.

5. CONCLUSIONS

In this paper we propose PMF-SVN, an efficient method to solve the tag completion problem. The key idea is to capture the semantical and visual correlations of images and to form SVNs, incorporating thus SVNs into the learning process of a probabilistic matrix factorization framework. Our experimental evaluation on benchmark datasets demonstrates the superiority of PMF-SVN over baselines. Our future work includes an adaptive strategy for both the SVN formation algorithm and the learning process of PMF to evaluate PMF-SVN on the evolving data of social media.

6. REFERENCES

[1] J. Chen, Y. Cui, G. Ye, D. Liu, and S. Chang. Event-driven semantic concept discovery by exploiting weakly tagged internet images. In *ICMR*, page 1, 2014.

[2] Z. Feng, S. Feng, R. Jin, and A. K. Jain. Image tag completion by noisy matrix recovery. In *ECCV*, pages 424–438, 2014.

[3] Z. Lin, G. Ding, M. Hu, J. Wang, and X. Ye. Image tag completion via image-specific and tag-specific linear sparse reconstructions. In *CVPR*, pages 1618–1625, 2013.

[4] B. Markines, C. Cattuto, F. Menczer, D. Benz, A. Hotho, and G. Stumme. Evaluating similarity measures for emergent semantics of social tagging. In *WWW*, pages 641–650, 2009.

[5] S. Moran and V. Lavrenko. Sparse kernel learning for image annotation. In *ICMR*, page 113, 2014.

[6] R. Salakhutdinov and A. Mnih. Probabilistic matrix factorization. In *NIPS*, pages 1257–1264, 2008.

[7] L. Wu, R. Jin, and A. K. Jain. Tag completion for image retrieval. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(3):716–727, 2013.

[8] G. Zhu, S. Yan, and Y. Ma. Image tag refinement towards low-rank, content-tag prior and error sparsity. In *ACM Multimedia*, pages 461–470, 2010.

[9] `http://lear.inrialpes.fr/people/guillaumin/data.php`.

[10] `https://sites.google.com/site/linzijia72/`.

[11] `http://www.cse.msu.edu/~fengzhey/downloads/src/`.