2012 IEEE International Conference on Multimedia and Expo

Finding Subgroups in a Flickr Group

Sumit Negi
IBM Research
Plot No.4, Block C, Institutional Area, Vasant Kunj
New Delhi, India
sumitneg@in.ibm.com

Santanu Chaudhury
Dept. of Electrical Engineering
Indian Institute of Technology, Delhi
New Delhi, India
santanuc@ee.iitd.ernet.in

***Abstract*—Information management systems today face a tremendous challenge considering the growing popularity of social media repositories involving images and video. Considering the growing volume of multimedia content in such online media-sharing communities there is an increasing need for novel ways of organizing content. In this paper we consider the problem of organizing images in a given *Flickr Group*[1] by discovering latent *subgroups*. A Flickr Group can be visualized as a collection of such subgroups where each subgroup represents a distinct theme. We model the task of discovering subgroups as that of finding highly correlated topics from a dataset containing images and associated tags. The proposed probabilistic model employs a more flexible prior distribution to model topic-topic correlations and utilizes both tag and image information for discovering such subgroups. Our experiments on Flickr Group data demonstrate that the model is able to successfully discover subgroups without any supervision.**

***Keywords*-Flickr, generative model, subgroup discovery task**

I. Introduction

Information management systems today face a tremendous challenge considering the growing popularity of social media repositories involving images and video. Falling prices of digital imaging devices and increased internet penetration has led to an exponential increase in the amount of multimedia content that is available online. A good example of this is seen in image-sharing online communities such as Flickr[2], Zoomr[3], PhotoBucket[4] etc. Flickr alone hosts over 5 billion user contributed images. Users, as part of such image-sharing communities, tag (*tagging is the practice of attaching short descriptive terms or keyword with images*) and share images with other users and also organize themselves into user-groups or communities around common interest. *Flickr groups* is such an example wherein members share images around a specific topic or theme (e.g. *San Francisco Bay*, *World Events (festivals, protests,etc.)*). The main purpose of groups is to facilitate the sharing of user photos in what is called the *group pool*[5]. This is a collection of photos shared by any member with the group, and, implicitly, all the tags associated with the photo become part of the group photo pool.

A Flickr group can thus be considered as a pool of photos contributed by group members around a specific theme. For instance, an actively administered user-group on "*18th century Anglo-Saxon architecture*" would contain images depicting architecture from this period. However, due to the broad nature of the topic it is likely that one would observe different *subgroups* or themes within this user-group. For example, one could expect images of 18th century Anglo-Saxon churches and forts to constitute two different *subgroups* or themes under the broad user-group on "*18th century Anglo-Saxon architecture*". Being able to discover such subgroups in an unsupervised manner from a user-group has its own merit. Such subgroups can help in automatically organizing content in a Flickr group thus aiding in it's navigation and maintenance.

Against this background, this paper focuses on the following problem: Given a dataset of images from a Flickr group and their associated captions (i.e user generated tags), can we build a model that automatically discovers subgroups using both image and tag features. To the best of our knowledge we are the first to pose this problem and propose an unsupervised method for discovering subgroups for a given Flickr group.

II. Prior Work

Flickr has been the object of many studies, including attempts to characterize users ([1], [2]), the tags users assign to photos ([3]), and investigating user's motivations for publishing and tagging ([4], [5]). Other studies have looked at how Flickr data can be used for a variety of purposes, such as providing recommendations for tagging photos ([6]), and automatically assigning geographic coordinates to Flickr photos ([7]) etc. Despite this there has been relatively little work on Flickr Groups as such. Exception to this is the work by [8], [9] and [10]. Authors in [8] propose a topic-based representation of flickr groups for characterizing and searching groups. This model was later extended to jointly represent both flickr users and groups [9]. In [10] the authors pose the problem of discovering *hypergroups* in Flickr,

[1]http://www.flickr.com/help/groups/
[2]http://www.flickr.com/
[3]http://www.zooomr.com/
[4]http://photobucket.com/
[5]http://www.flickr.com/help/groups/

978-0-7695-4711-4/12 $26.00 © 2012 IEEE
DOI 10.1109/ICME.2012.114

i.e., communities consisting of collection of Flickr groups. Our work differs from previous work in the following two respects.

- To the best of our knowledge we are the first to pose the problem of *subgroup* discovery in a given Flickr group. Unlike *hypergroups* [10], which are communities consisting of collection of Flickr groups, we seek to partition an individual Flickr group into subgroups where each subgroup represents a distinct theme under that group.
- Our proposed generative model uses both image and tag information for discovering such subgroups.

III. Approach

Our model is inspired by the Multi-modal Latent Dirichlet Allocation (MoM-LDA) [11] work (also referred to as *correspondence latent dirichlet allocation*) which looks at the problem of modeling annotated data – data with multiple types where the instance of one type (such as a caption) serves as a description of the other type (such as an image). MoM-LDA is an extension of the popular Latent Dirichlet Allocation (LDA) [12] model for modeling collections of discrete data such as text corpora. In topic modeling for text documents LDA assumes the following generative process: each document has its own distribution of topics, and given a specific topic, the words are generated. MoM-LDA is a generalization of LDA where the documents contain multiple types (modalities) of entities such as words, image regions (also called blobs). MoM-LDA describes the following generative process for the data: each document (consisting of both words and pictures) has a distribution for a fixed number of mixture components (topics), and given a specific mixture component the words and the image features are generated. We further extend the MoM-LDA model to allow for

- Topic correlations: The existing MoM-LDA model does not consider correlation between topics. Like LDA, MoM-LDA assumes that the discovered topics are independent of each other. This limitation stems from the independence assumptions implicit in the Dirichlet distribution on the topic proportions. Under a Dirichlet, the components of the proportions vector are nearly independent; this leads to the strong and unrealistic modeling assumption that the presence of one topic is not correlated with the presence of another. In any corpus it is likely that subsets of the underlying latent topics will be highly correlated. For instance, image and tag features for the '*bridge*' (a structure built to span physical obstacles) topic would be highly correlated with image and tag feature for the '*water*' topic. We exploit this correlation to discover subgroup structures. Topics that are strongly correlated with one another are treated as belonging to one subgroup. To model topic correlations we utilize a more flexible distribution for the topic proportions namely the *logistic normal distribution*. The logistic normal models correlations between components of the random variable through the covariance matrix of the normal distribution. Correlated Topic Model [13] uses a similar prior for discovering correlated topics from text documents. To the best of our knowledge we are the first to apply this prior in the context of tagged images.
- Visual Words: In our model we use a visual-word [14] image representation which is analogous to the bag-of-words representation for text documents. The MoM-LDA model as described in [11] assumes that the image region are generated from a multivariate Gaussian distribution. To incorporate the visual-word generation of image regions we modify the MoM-LDA model to use a multinomial distribution instead of the multivariate Gaussian distribution. This is motivated by [15].

The resulting model is shown in Figure 1. The generative process is as follows

1) Draw $\eta \sim$ N(μ,Σ) (where N(μ,Σ) is a K dimensional Normal distribution).
2) For each visual word v_m, $m \in \{1,.....,D_M\}$.
   a) Draw topic assignment $z_m \mid \eta \sim$ Multinomial($f(\eta)$).
   b) Draw visual word $v_m \mid z_m \sim$ Multinomial(π_{z_m}).
3) For each tag word w_n, $n \in \{1,.....,D_N\}$.
   a) Draw indexing variable $y_n \sim$ Unif(1,.....D_M)
   b) Draw tag word $w_n \sim$ Multinomial($\beta_{z_{y_n}}$)

Where f(η_i)=$\frac{exp\ \eta_i}{\sum_j^K exp\ \eta_j}$ and D_M and D_N represents the number of visual and tag words associated with an image D respectively. In Figure 1 nodes represent random variables and edges indicate possible dependence. Shaded nodes are observed random variables; unshaded nodes are latent random variables. The box around a random variable is a plate,

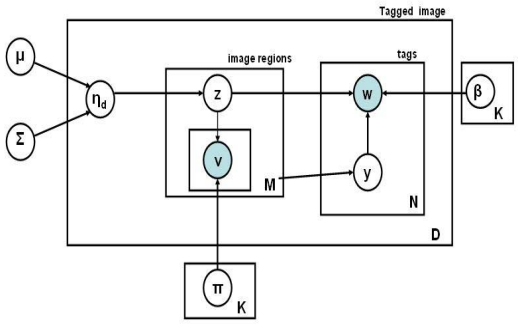

Figure 1. Probabilistic model for subgroup discovery

a notational device to denote replication. The box around w denotes N replicates of w. Table I provides details of the

symbols used in the plate model.

Table I
SYMBOLS USED IN THE MODEL

Symbol	Description
M	Size of visual-word vocabulary
N	Size of tag-word vocabulary
D	Number of images with tags
K	Number of topics (user-specified)
z	Latent factor that generates images
y	Latent factor that generates tags
v	Observed visual word
w	Observed tag word
β	Multinomial over tag-word vocabulary
π	Multinomial over visual-word vocabulary
Σ	K×K dimensional covariance matrix
μ	K dimensional mean vector
η	Draw from a gaussian distribution N(μ,Σ)

The proposed model is used to estimate the topic-topic correlation (i.e. the estimated value of the K $\times$ K covariance matrix Σ, where K is the number of topics) for a given Flickr group (i.e. using the tagged image dataset corresponding to that Flickr group). The estimated covariance matrix is then used to partition the discovered topics into separate subgroups based on the correlation that exists between topics. To partition topics into subgroups all entries in the covariance matrix that are less than a user specified threshold are set to 0 (i.e. weak topic-topic correlations are ignored). Topics with high correlation amongst/between themselves are treated as part of a separate subgroup. This is illustrated in Figure 2 where Topic 1 is strongly correlated with Topic 2 and Topic 4 (this is depicted using $\times$ in the matrix). Consequently these three topics form a subgroup (Note: Topic 2 is connected to Topic 4 via Topic 1). Similarly, Topic 3 and Topic 5 form their own subgroup (on account of being highly correlated). Note that the empty cells in Figure 2 denote entries in the matrix which have been set to 0 on account of the fact that the correlation value is less than a user specified threshold. Please note all diagonal entries in the covariance matrix are ignored for the subgroup formation process.

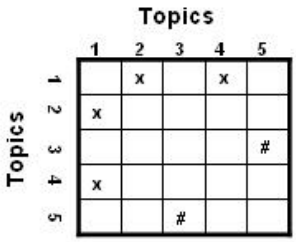

Figure 2. Partitioning topics into subgroups using covariance matrix

A. Variational Inference

The joint distribution of the model in Figure 1 is given by

$$\begin{aligned} &p(\mathbf{v}, \mathbf{w}, \eta, \mathbf{z}, \mathbf{y}|\mu, \Sigma, \pi, \beta) \\ &= p(\eta|\mu, \Sigma) \times \prod_{m=1}^{M} p(z_m|\eta) \times p(v_m|z_m, \pi) \\ &\prod_{n=1}^{N} p(y_n|M) \times p(w_n|y_n, \mathbf{z}, \beta) \end{aligned} \tag{1}$$

where μ, Σ, π, β are the parameters we want to estimate. Boldfaced fonts refer to vector. The **likelihood** is written by marginalizing over the hidden variables η, z, y.

$$\begin{aligned} &p(\mathbf{v}, \mathbf{w}|\mu, \Sigma, \pi, \beta) \\ &= \int_{\eta} \sum_{z} \sum_{y} p(\mathbf{v}, \mathbf{w}, \eta, \mathbf{z}, \mathbf{y}|\mu, \Sigma, \pi, \beta) d\eta \end{aligned} \tag{2}$$

This integral is intractable due to the coupling of η and π. We therefore resort to variational inferencing to approximate the posterior distribution over latent variables given a image and it's tags. The original graphical model is simplified and free variational parameters introduced [15]. Equation (3) shows the simplified posterior distribution and it's factorization.

$$\begin{aligned} &q(\eta_{1:K}, \mathbf{z}, \mathbf{y}|\alpha_{1:K}, \nu^2_{1:K}, \phi, \lambda) \\ &= \prod_{i=1}^{K} q(\eta_i|\alpha_i, \nu_i^2) \times \prod_{m=1}^{M} q(z_m|\phi_m) \times \prod_{n=1}^{N} q(y_n|\lambda_n) \end{aligned} \tag{3}$$

where each $q(\eta_i|\alpha_i, \nu_i^2)$ is a univariate gaussian with parameters α_i, ν_i^2 and $q(z_m|\phi_m)$, $q(y_n|\lambda_n)$ are multinomial with parameters ϕ_m and λ_n respectively.

The log-likelihood can now be written as

$$\begin{aligned} &log\ p(\mathbf{v}, \mathbf{w}|\mu, \Sigma, \pi, \beta) \\ &= log \int_{\eta} \sum_{z} \sum_{y} \frac{p(\mathbf{v}, \mathbf{w}, \eta, \mathbf{z}, \mathbf{y}|\mu, \Sigma, \pi, \beta) q(\eta_{1:K}, \mathbf{z}, \mathbf{y}|\alpha_{1:K}, \nu^2_{1:K}, \phi, \lambda)}{q(\eta_{1:K}, \mathbf{z}, \mathbf{y}|\alpha_{1:K}, \nu^2_{1:K}, \phi, \lambda)} d\eta \end{aligned} \tag{4}$$

Applying Jensen's inequality the log-likelihood is written as

$$\begin{aligned} &log\ p(\mathbf{v}, \mathbf{w}|\mu, \Sigma, \pi, \beta) \\ &\geqslant E_q[log\ p(\mathbf{v}, \mathbf{w}, \eta, \mathbf{z}, \mathbf{y}|\mu, \Sigma, \pi, \beta)] \\ &- E_q[log\ q(\eta_{1:K}, \mathbf{z}, \mathbf{y}|\alpha_{1:K}, \nu^2_{1:K}, \phi, \lambda)] \end{aligned} \tag{5}$$

The RHS of equation (5) gives the lower bound on the log-likelihood of a tagged image and is refereed to as $L(\alpha_{1:K}, \nu^2_{1:K}, \phi, \lambda; \mu, \Sigma, \pi, \beta)$ where E_q denote expectation over the probability distribution q i.e. equation (3). By using equation (1) and (3) we can write (5) as

$$\begin{aligned} &L(\alpha_{1:K}, \nu^2_{1:K}, \phi, \lambda; \mu, \Sigma, \pi, \beta) \\ &= E_q[log\ p(\eta|\mu, \Sigma)] + E_q[log\ p(\mathbf{z}|\eta)] + E_q[log\ p(\mathbf{v}|\mathbf{z}, \pi)] \\ &+ E_q[log\ p(\mathbf{y}|M)] + E_q[log\ p(\mathbf{w}|\mathbf{y}, \mathbf{z}, \beta)] \\ &- E_q[log\ q(\eta_{1:K}|\alpha_{1:K}, \nu^2_{1:K})] \\ &- E_q[log\ q(\mathbf{z}|\phi)] - E_q[log\ q(\mathbf{y}|\lambda)] \end{aligned} \tag{6}$$

The first two terms of equation (6) are similar to those in the Correlated-LDA model [13] namely $E_q[log\ p(\eta|\mu,\Sigma)]$ and $E_q[log\ p(\mathbf{z}|\eta)]$. Readers are requested to refer to [13] for their expansion.

The other $E_q[..]$ terms in equation (6) can be written as

$$E_q[log\ p(\mathbf{v}|\mathbf{z},\pi)] = \sum_{m=1}^{D_M}\sum_{i=1}^{K}\sum_{j=1}^{M} E_q[z_m^i]v_m^j log\ \pi_{ij} \quad (7)$$

$$E_q[log\ p(\mathbf{w}|\mathbf{y},\mathbf{z},\beta)] = \sum_{n=1}^{D_N}\sum_{i=1}^{K}\sum_{j=1}^{N}\sum_{m=1}^{D_M} \phi_{mi}\lambda_{nm}w_n^j\ log\ \beta_{ij} \quad (8)$$

$$E_q[log\ q(\eta_{1:K}|\alpha_{1:K},\nu^2_{1:K})] = \sum_{i=1}^{K} E_q[q(\eta_i|\alpha_i,\nu_i^2)] \quad (9)$$

$$E_q[log\ q(\mathbf{z}|\phi)] = \sum_{m=1}^{D_M}\sum_{i=1}^{K} \phi_{mi}\ log\ \phi_{mi} \quad (10)$$

$$E_q[log\ q(\mathbf{y}|\lambda)] = \sum_{n=1}^{D_N}\sum_{m=1}^{D_M} \lambda_{nm}\ log\ \lambda_{nm} \quad (11)$$

where – D_M and D_N represents the number of visual and tag words associated with an image D respectively, v_m^j and w_n^j are indicator variables which are set to 1 if v_m (or w_n) is the same as the j^{th} token in the visual (or tag) vocabulary else these are set to 0, similarly z_m^i = 1 if z_m is the i^{th} latent topic, else it is 0. Equation 7-11 are put back into equation (6) and the resulting equation is maximized w.r.t to the variational parameters $\alpha_{1:K}$,$\nu^2_{1:K}$, ϕ and λ.

Having maximized the lower bound (E-step) w.r.t. the variational parameters we next maximize the bound w.r.t. to the model parameters Σ, π, β (M-step). The M-step involves calculating the maximum likelihood estimation of the topics and multivariate Gaussian using expected sufficient statistics [16], where the expectation is taken with respect to the variational distributions computed in the E-step. The E-step and M-step are repeated until the bound on the likelihood converges.

B. Visual Features

This section describes the method used to build the visual vocabulary [14]. Each image in the data-set is represented as a *bag-of-visual* words. First, keypoints or local interest points are detected for each image. Keypoints are salient image patches that contain rich local information about an image, which can be automatically detected using various detectors ([17],[18],[19]). These keypoints are represented using one or more local descriptors (e.g. SIFT, texture, color etc). Keypoints are then grouped (using k-means) into a large number of clusters with those with similar descriptors assigned into the same cluster. By treating each cluster as a "*visual word*" that represents the specific local pattern shared by all the keypoints in that cluster, a visual-word vocabulary describing all kinds of local image patterns is built.

IV. Experiment

The objective of our experiment is to demonstrate (both through specific examples and empirically) that the proposed model is able to accurately discover coherent subgroups for a given Flickr group.

A. Data-Set

We test our model on 5 different datasets. Each dataset is prepared by crawling images and associated metadata (contributing member name, tags) from specific Flickr group. Details of the Flickr group used and the number of images collected from that group is shown in Table II. To ensure diversity we retain only a maximum of 80 images per member in a given dataset. Moreover, only those images which have 3 or more tag words associated with it (after removal of tags that contained numeric characters e.g. dates or years) were retained. For the data collection task Flickr APIs[6] were used.

Table II
DATASET DESCRIPTION

Dataset Name	Flickr Group Name	Dataset Size
A	*AVIATION "MACRO" PHOTOGRAPHY*	2100
B	*Architecture in Europe*	2300
C	*Monuments*	1800
D	*Gardens To See (NO FLOWER CLOSE-UPS, PLEASE!)*	2170
E	*Architecture Photography*	2000

B. Visual and Tag word vocabulary

For our experiments we use a Harris-Laplace operator to detect keypoints or interest points (IP) in images. For each keypoint/IP a local descriptor is calculated using four appearance features − color, texture, location and quantized SIFT. All the local descriptors are then grouped together into clusters using k-means. For all our datasets we use a visual-vocabulary of size 300 and a tag-vocabulary of size 240. Please note that each dataset has it's own unique visual and tag vocabulary and hence the visual and tag vocabulary is generated individually for each dataset.

C. Discovered Subgroups

To give readers an idea of the kind of subgroups that were obtained by our model we show the subgroups for Dataset *A* and *D* in Figure 3. K i.e. the number of topics was set to 14 and 10 for the two groups respectively (i.e. $K_A = 14$ and $K_D = 10$). We fit the model described in Section III to the two datasets separately and obtain estimates of the respective covariance matrices (Σ_A, Σ_D). These matrices are then used to partition the topics under each datasets into subgroups using the approach outlined in Section III. For instance, in Dataset *A* the 14 topics [$T1$,$T2$,...,$T14$] are partitioned into 3 subgroups namely $C1$=[$T1$,$T4$,$T7$], $C2$=[$T3$,$T10$] and $C3$=[$T8$,$T9$,$T12$]. This is shown in Figure 3, where the presence (or absence) of edges between topics denotes strong (or weak/absent) correlation. With each topic we display the top five tag words associated with that topic (e.g. $T1$=["*f16*", "*f18*", "*f-15e*", "*fighter*", "*military*"]. One can observe from Figure 3 that the discovered subgroups (namely $C1$,$C2$ and $C3$) divide the Dataset *A* into distinct themes - where $C1$ predominantly containing images and tags referring to "*commercial airlines*", $C2$ "*military aircrafts*" and $C3$ "*engine parts*". Similarly, the discovered subgroups in Dataset *D* (Figure 3) divide the dataset into distinct themes - where $C1$ mainly contains images and tags referring to "*palaces and buildings*", $C2$ "*gardens*" and $C3$ "*water bodies and bridges*".

[6]http://www.flickr.com/services/api/

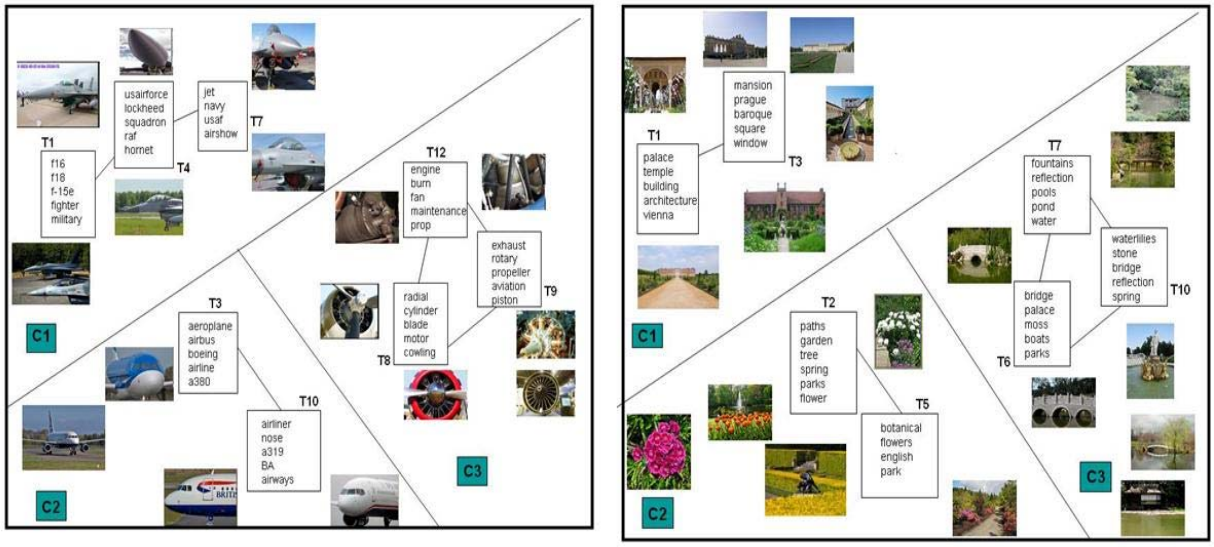

Figure 3. Discovered subgroups for Dataset *A* [left] and Dataset *D* [right] {best viewed in color}

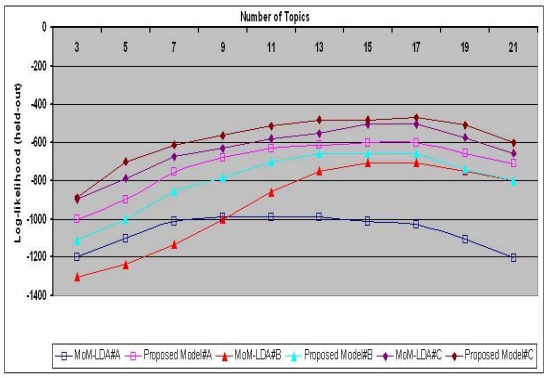

Figure 4. Held-out log-likelihood for Datasets *A*, *B* and *C*

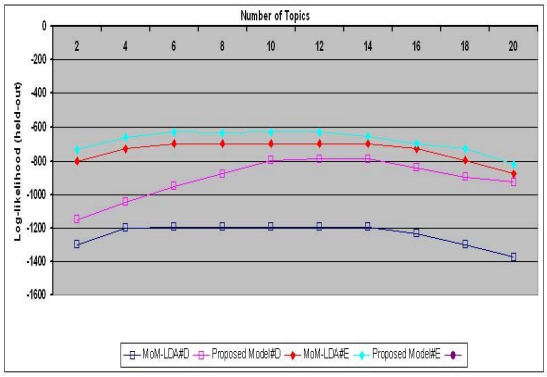

Figure 5. Held-out log-likelihood for Datasets *D* and *E*

D. Empirical Results : Perplexity Analysis

For an empirical comparison we compare our model to MoM-LDA (which also uses text and image features) by fitting a smaller collection of images to models of varying numbers of topics. For each dataset 15% of the images are used as the test (held-out) dataset. Using ten-fold cross validation, we compute the log-likelihood for each test dataset separately (the corresponding model parameters are estimated using 85% of the dataset). A better model of the image collection will assign higher log-likelihood to the held-out dataset. Figure 4 and Figure 5 compares the average held out log likelihood for both the MoM-LDA and our proposed model on all the 5 datasets. Please note the legend *MoM-LDA#A* in Figure 4 indicates log-likelihood results for MoM-LDA on dataset A, similarly legend *Proposed Model#A* indicates log-likelihood results for the proposed model on dataset A. Consistent high value of log-likelihood across the entire topic range indicates that our model fits the data better than MoM-LDA.

V. Conclusion and Future Work

In this paper we considered the problem of discovering *subgroups* in a given *Flickr Group*. A Flickr Group can be visualized as a collection of such subgroups where each subgroup represents a distinct theme. We model the task of discovering subgroups as that of finding highly correlated topics from a dataset containing images and associated tags. The proposed probabilistic model employs a more flexible prior distribution to model topic-topic correlations and utilizes both tag and image information for discovering subgroups. Our experiments on Flickr Group data demonstrate that the model is able to successfully discover subgroups from a Flickr group without any supervision. As part of future work we would like to investigate the non-parametric variant of our model. One limitation of parametric Bayesian models (such as LDA) is that one needs to specify the number of components of the mixture model. In practice, the model is trained for several choices of the mixture components and the optimal number of components is chosen based on the validation on a held-out set. This can be expensive because of the need to train several models. Our model also suffers from such a limitation (one needs to specify the number of topics i.e. K). To address this limitation non-parametric generalization of some popular models have been proposed. These non-parametric models such as the hierarchical Dirichlet Process learn the number of components directly from the data. For instance, the MoM-HDP model [20] is an example of a non-parametric generalization of

the MoM-LDA model which uses a hierarchical Dirichlet Process. Obtaining a corresponding non-parametric generalization for our model is difficult due to the presence of a non-exponential family prior. We plan to investigate such a generalization as part of our future work.

REFERENCES

[1] A. Miller and W. Edwards, "Give and take: A study of consumer photo-sharing culture and practice," *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 347–356, 2007.

[2] A. Cox, P. Clough, and J. Marlow, "Flickr: A first look at user behaviorin the context of photography as serious leisure," *Information Research*, 2008.

[3] A. Rorissa, "A comparative study of flickr tags and index terms in a general image collection," *Journal of theAmerican Society for Information Science and Technology*, pp. 2230–2242, 2010.

[4] O. Nov, M. Naaman, and C. Ye, "What drives content tagging: The case of photos on flickr?" *In Proceedings of the 26th Annual SIGCHI Conference on Human Factors in Computing Systems*, pp. 1097–1100, 2008.

[5] E. Angus, M. Thelwall, and D. Stuart, "General patterns of tag usage among university groups in flickr," *Online Information Review*, pp. 89–101, 2008.

[6] B. Sigurbjrnsson and R. V. Zwol, "Flickr tag recommendation based on collective knowledge," *In Proceedings of the 17th international Conference on World Wide Web*, pp. 327–336, 2008.

[7] O. V. Laere, S. Schockaert, and B. Dhoedt, "Towards automated georeferencing of flickr photos," *Proceedings of the Sixth Workshop on Geographic information Retrieval*, pp. 1–7, 2010.

[8] R. Negoescu and D. Gatica-Perez, "Analyzing flickr groups," *Proceedings of the 2008 International Conference on Content-Based Image and Video Retrieval*, pp. 417–426, 2008.

[9] R. A. Negoescu and D. Gatica-Perez, "Topickr: Flickr groups and users reloaded," *In Proceedings of the 16th ACM International Conference on Multimedia*, 2008.

[10] R. Negoescu, B. Adams, D. Phung, S. Venkatesh, and D. Gatica-Perez, "Flickr hypergroups," *In Proceedings of the 17th ACM International Conference on Multimedia*, 2009.

[11] D. M. Blei and M. I. Jordan, "Modeling annotated data," *In Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pp. 127–134, 2003.

[12] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research*, p. 9931022, 2003.

[13] D. M. Blei and J. D. Lafferty, "Correlated topic models," *Advances in Neural Information Processing Systems*, 2006.

[14] J. Yang, Y.-G. Jiang, A. G. Hauptmann, and C.-W. Ngo, "Evaluating bag-of-visual-words representations in scene classification," *In Proceedings of the international workshop on Multimedia Information Retrieval*, pp. 197–206, 2007.

[15] H. Xiao, "Derivation of variational inference for correspondence-lda model," *Tech Report*, 2010.

[16] P. Hoff, "Nonparametric modelling of hierarchically exchangeable data," *Tech Report*, 2003.

[17] C. Harris and M. Stephens, "A combined corner and edge detector," *In Alvey Vision Conference*, pp. 147–151, 1988.

[18] A. Heyden and K. Rohr, "Evaluation of corner extraction schemes using invariance methods," *In Proceedings of the 13th International Conference on Pattern Recognition*, pp. 895–899, 1996.

[19] F. Mokhtarian and R. Suomela, "Robust image corner detection through curvature scale space," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, p. 13761381, 1998.

[20] O. Yakhnenko and V. Honavar, "Multi-modal hierarchical dirichlet process model for predicting image annotation and image-object label correspondence," *In SIAM SDM*, 2009.