PLSA DRIVEN IMAGE ANNOTATION, CLASSIFICATION, AND TOURISM RECOMMENDATION

Konstantinos Pliakos and Constantine Kotropoulos

Department of Informatics, Aristotle University of Thessaloniki
Thessaloniki, 54124, Greece
Email: {kpliakos, costas}@aiia.csd.auth.gr

ABSTRACT

A burst of interest in image annotation and recommendation has been witnessed. Despite the huge effort made by the scientific community in the aforementioned research areas, accuracy or efficiency still remain open problems. Here, efficient methods for image annotation, visual image content classification as well as touristic place of interest (POI) recommendation are developed within the same framework. In particular, semantic image annotation and touristic POI recommendation harness the geo-information associated to images. Both semantic image annotation and visual image content classification resort to Probabilistic Latent Semantic Analysis (PLSA). Several tourist destinations, strongly related to the query image, are recommended, using hypergraph ranking. Experimental results were conducted on a large image dataset of Greek sites, demonstrating the potential of the proposed methods. Semantic image annotation by means of PLSA has achieved an average precision of 90% at 10% recall. The average accuracy of content-based image classification is 80%. An average precision of 90% is measured at 1% recall for tourism recommendation.

***Index Terms*—** Probabilistic Latent Semantic Analysis (PLSA), Image Classification, Image Annotation, Recommender systems, Hypergraph

1. INTRODUCTION

Nowadays, the deployment of many photo-sharing web applications with rising popularity has increased tremendously the amount of images uploaded to the web. Consequently issues related to search and organization have emerged, amplifying the need for efficient annotation and recommendation algorithms. Several websites like *Flickr*[1] or *Picasa Web Album*[2] enable users to annotate images, describing their content. Image annotation aims at bridging the semantic gap between the semantic and visual content of an image. Furthermore, it affects significantly the retrieval accuracy of search engines, which are based heavily on the text information provided with images, such as tags, titles, etc.

During the last years, besides annotation, increasing interest in efficient recommendation has been witnessed. Indeed, brochures or simple web search have been substituted by tourist recommendation systems. Despite the effort that has been made so far, there are open problems in accuracy and efficiency to be addressed.

In the past, many efforts were made toward image annotation. In [1], an image and video annotation model was proposed based on the joint probability distribution of tags and image feature vectors. The tag probabilities were computed using a multiple Bernoulli model and the probabilities of image features were obtained using non-parametric kernel density estimates. A joint probabilistic model was proposed for simultaneous image classification and annotation in [2]. It was based on a multi-class extension of the supervised Latent Dirichlet Analysis (sLDA) [3]. Graph-based methods were proposed in [4, 5] for tag recommendation, capturing the information from multi-type interrelated objects. A related work in tourism recommendation is that of L. Cao *et al.* [6], where recommendation was based on clustering of geotagged images by location and visual matching.

The main contribution of this paper is in the development of efficient semantic image annotation, content-based image classification, and tourism recommendation methods in a unifying canvas. Indeed, semantic image annotation and tourism recommendation harness the geo-tag information of images. Probabilistic Latent Semantic Analysis (PLSA) is the heart of the methods for semantic image annotation and content-based image classification, and pertains the hypergraph ranking employed for tourism recommendation. In addition, the proposed content-based image classification is exploited to propagate labels associated with visual content classes to images, enhancing further the semantic image annotation.

To begin with, geo-tagged images are first clustered by location (latitude, longitude), forming several geographical clusters, called $geo-clusters$ hereafter. The geo-clusters are then sorted with respect to their density (i.e., the number of images they contain) to define the places of interest (POIs) to tourists. The underlying rationale is that popular tourist destinations attract more visitors, who upload more geo-tagged images on the web. For each geo-cluster, a document is formed by concatenating the text information (e.g., title, tags) associated to the images that belong to this geo-cluster. Next, a term-document matrix is created and PLSA [7–9] is applied to it. Replacing PLSA with LDA [10] does not improve the annotation performance. Thus, PLSA is preferred due to its simplicity. During the annotation, the most strongly related terms to the prevailing topic of each geo-cluster are assigned to it and propagated to all its constituent images.

Semantic image annotation is complemented by content-based image classification based on the PLSA applied to the visual word-image matrix. Both SIFT [11] and GIST [12] descriptors are exploited to classify each image to a predefined number of classes associated to a large image dataset. The class label is treated as a complementary image description.

Tourism recommendation is based on a hypergraph [13] whose vertices are the annotation terms, the geo-cluster documents, and the latent topics derived by the PLSA. The hyperedges of the hypergraph

[1] http://www.flickr.com
[2] http://picasaweb.google.com

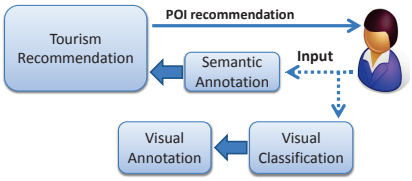

Fig. 1. Annotation and recommendation system.

capture the high-order relations between the vertices in contrast to the edges of common graphs. Tourism recommendation is treated as a hypergraph ranking problem, recommending the top ranked geo-clusters as touristic destinations.

The block diagram of the proposed methods is depicted in Fig. 1. The user gives a test image as input to the system. The image is assigned to a geo-cluster according to its GPS coordinates and is annotated semantically and geographically, as it is described in Section 3.1. Simultaneously, the image is classified visually into one of a predefined number of classes, as it is detailed in Section 3.2. The class label and its associated representative tags are exploited for visual content annotation. Proceeding to tourism recommendation, the query vector is set, as in Section 4.1. Hypergraph ranking is applied to geo-cluster documents, topics, and terms and the top ranked geo-cluster documents are recommended as touristic POIs.

Promising experimental results are disclosed. In particular, an average precision of 90% at 10% recall is reported for semantic image annotation. The average accuracy of content-based image classification of 205 test images over 13 classes is 80%, when both SIFT and GIST features are exploited. For tourism recommendation, an average precision of 90% is measured at 1% recall, indicating the effectiveness of the proposed recommendation method.

The remainder of the paper is organized as follows. In Section 2 the dataset description is presented. In Section 3, semantic image annotation and visual image classification and annotation are detailed. The hypergraph ranking model and the hypergraph construction are described in Section 4. In Section 5, experimental results are reported, demonstrating the effectiveness of the proposed method. Conclusions are drawn and topics for future research are indicated in Section 6.

2. DATASET

A dataset of 50000 images related to Greek sites was collected from $Flickr$. Sample images are depicted in Fig. 2. The geo-tags (GPS coordinates) of these images were clustered into 4660 clusters by means of hierarchical clustering applied to distances computed using the "Haversine formula"[3]. From these geo-clusters, only the 2000 most dense were considered as touristic (POIs), containing 45316 images. For each geo-cluster, a document was created by concatenating the text information (e.g., title, tags) of all its images.

Next, text information related to 150000 images was crawled in order to properly capture the context of the tourism application. All characters were converted to lower case. Unreadable symbols and redundant information were removed. Terms with frequency less than 100 were eliminated, yielding a vocabulary of 1901 terms.

[3]http://www.movable-type.co.uk/scripts/latlong.html

Fig. 2. A sample of 16 images from the dataset.

3. IMAGE ANNOTATION

3.1. Image Annotation Using Semantic Topics

PLSA performs a probabilistic mixture decomposition, which associates an unobserved class variable to co-occurrences of terms and documents. By applying PLSA to a term-document matrix, the relations between the terms and the documents are captured by the probability distribution between the documents and the generated topics as well as between the topics and the terms. PLSA models each term in a document as a sample from a mixture model. The mixture components are multinomial random variables that can be interpreted as topic representations. The data generation process can be described as follows, [7, 9]: 1) select a document d with probability $P(d)$, 2) pick a latent topic z_a with probability $P(z_a|d)$ and, 3) generate term t_a with probability $P(t_a|z_a)$.

Let $t_a \in T_a = \{t_{a_1}, t_{a_2}, \cdots, t_{a_k}\}$ be a vocabulary term and $d \in D = \{d_1, d_2, \cdots, d_m\}$ denote a document. The joint probability model is defined by the mixture:

$$\left.\begin{aligned} P(t_a, d) &= P(d)P(t_a|d) \\ P(t_a|d) &= \sum_{z_a \in Z_a} P(t_a|z_a)P(z_a|d) \end{aligned}\right\} \quad (1)$$

where $z_a \in Z_a = \{z_{a_1}, z_{a_2}, \cdots, z_{a_n}\}$ is an unobserved class variable representing the topics. As it is indicated in (1), the document specific term distribution $P(t_a|d)$ is a convex combination of the n topic dependent distributions $P(t_a|z_a)$. The annotation procedure is performed as follows:

1 PLSA is applied to a term-document matrix $\mathbf{A} \in \mathbb{R}^{k \times m}$. Here, the documents are formed by concatenating any terms in the tags or the title of the images that belong to a geo-cluster. Any document $d \in D$ is represented by a vector of size k, having as elements the frequency of occurrence of each term in d.

2 For each document to be annotated, the most related topic is chosen, that with the highest probability, i.e., $z_a^* = \arg\max_{z_a \in Z_a} P(z_a|d)$.

3 The $k' << k$ most related terms to z_a^* are identified by sorting $P(t_a|z_a^*)$ in decreasing order of magnitude.

Here, the term document matrix $\mathbf{A}$ is of size 1901×2000. Among the most descriptive terms of a document, those providing geographical information are identified using geo-gazetteers[4]. Thus,

[4]http://www.geonames.org

a complete annotation model is built, which provides geographic information in addition to the semantic information.

3.2. Visual Classification and Annotation

The semantic annotation, detailed in Section 3.1, is complemented by visual annotation based on scene classification. Scale-invariant feature transform (SIFT) [11] and GIST [12] descriptors are extracted from any image. Different visual classes $c \in C = \{c_1, c_2, \cdots, c_p\}$ have been defined, capturing the different themes pertaining the image dataset. The objective is to propagate the class label along with the associated tags to each image as visual annotation.

To construct a proper visual word vocabulary, a small image subset $G = \{g_1, g_2, \cdots, g_\nu\}$, made of images without occlusion or unwanted artifacts, is manually extracted and annotated using the p class labels. K means is applied to the SIFT descriptors of any image in the controlled dataset G, in order to quantize them to a predefined number (e.g., 200) of cluster mean vectors as codevectors. Any image $g \in G$ is represented by the concatenation of the codevectors $\tilde{g}$, instead of the concatenation of SIFT descriptors. K means is applied to the set of the aforementioned reduced representation $\tilde{g}$ in order to create the visual word vocabulary. The indices of the resulting codevectors are treated as visual words $t_v \in T_v = \{t_{v_1}, t_{v_2}, \cdots, t_{v_\kappa}\}$, where κ is the size of visual word vocabulary. Let $\mathbf{B} \in \mathbb{R}^{\kappa \times \nu}$ be the visual word-image matrix, having as columns the image representations built by measuring the frequency of visual words the reduced representations are quantized into. Similar to [14], PLSA is applied to $\mathbf{B}$ in order to calculate the conditional distributions $P(t_v|z_v)$ and $P(z_v|\tilde{g})$, where $z_v \in Z_v = \{z_{v_1}, z_{v_2}, \cdots, z_{v_l}\}$ are the visual latent topics.

Having learned the aforementioned conditional distributions from G, any test image g_{test} is represented by the conditional distribution $P(z_v|g_{test})$, obtained by running the M step of the Expectation Maximization (EM) algorithm for $P(z_v|g_{test})$ until convergence, keeping $P(t_v|z_v)$ fixed to these learned during the training.

Next, the κ_G nearest neighbors of the GIST descriptor extracted from any test image g_{test} are identified, using the K-nearest neighbor (KNN) classifier, which employs the Euclidean distances between g_{test} and any image in the controlled subset $g \in G$. Let $G_{NN}(g_{test})$ be the set of nearest neighbors. $G_{NN}(g_{test})$ is further narrowed to $\kappa_{GR} << \kappa_G$ training images by sorting the χ^2 distances between $P(z_v|g_{test})$ and $P(z_v|g)$, retaining the images associated to the κ_{GR} smallest distances. Let $\tilde{G}_{NN}(g_{test})$ be the resulting narrow set. Finally, the test image is assigned to the visual class c being in majority within the narrow set $\tilde{G}_{NN}(g_{test})$.

4. TOURISM RECOMMENDATION

A hypergraph is created to capture the multi-link relations between the vocabulary terms t_a, the geo-clusters d, and the topics z_a, computed in Section 3.1. Hereafter, set cardinality is denoted by $|\cdot|$, the ℓ_2 norm of a vector appears as $\|.\|_2$ and $\mathbf{I}$ is the identity matrix of compatible dimensions. $\Psi(V, E, w)$ denotes a hypergraph $\mathbf{H}$, with set of vertices V and set of hyperedges E to which a weight function $w : E \to \mathbb{R}$ is assigned. V consists of objects of different type (geo-clusters, topics, terms). An incidence matrix $\mathbf{H}$ of size $|V| \times |E|$ is formed, having elements $H(v, e) = 1$, if $v \in e$ and 0 otherwise. The vertex and hyperedge degrees are then defined as:

$$\left.\begin{aligned} \delta(v) &= \textstyle\sum_{e \in E} w(e) H(v, e) \\ \delta(e) &= \textstyle\sum_{v \in V} H(v, e) \end{aligned}\right\}. \tag{2}$$

The following diagonal matrices are defined: the vertex degree matrix $\mathbf{D}_u$ of size $|V| \times |V|$, the hyperedge degree matrix $\mathbf{D}_e$ of size $|E| \times |E|$, and the $|E| \times |E|$ matrix $\mathbf{W}$, containing the hyperedge weights defined in Section 4.1.

Let $\mathbf{\Theta} = \mathbf{D}_u^{-1/2}\mathbf{H}\mathbf{W}\mathbf{D}_e^{-1}\mathbf{H}^T\mathbf{D}_u^{-1/2}$. Then, $\mathbf{L} = \mathbf{I} - \mathbf{\Theta}$ is the positive semi-definite Laplacian matrix of the hypergraph. The elements of $\mathbf{\Theta}$, $\Theta(j, i)$, indicate the relatedness between the j and i. For clustering, a real-valued ranking vector $\mathbf{f} \in \mathbb{R}^{|V|}$ is sought that minimizes $\Omega(\mathbf{f}) = \frac{1}{2}\mathbf{f}^T\mathbf{L}\mathbf{f}$, requiring all vertices with the same value in the ranking vector $\mathbf{f}$ to be strongly connected [15]. The aforementioned optimization problem was extended by including the ℓ_2 regularization norm between the ranking vector $\mathbf{f}$ and the query vector $\mathbf{y} \in \mathbb{R}^{|V|}$ [16]. The function to be minimized is then expressed as

$$\tilde{Q}(\mathbf{f}) = \Omega(\mathbf{f}) + \vartheta \, ||\mathbf{f} - \mathbf{y}||_2^2 \tag{3}$$

where ϑ is a regularizing parameter. The best ranking vector, $\mathbf{f}^* = \arg\min_{\mathbf{f}} \tilde{Q}(\mathbf{f})$, is [16]:

$$\mathbf{f}^* = \frac{\vartheta}{1+\vartheta}\left(\mathbf{I} - \frac{1}{1+\vartheta}\mathbf{\Theta}\right)^{-1}\mathbf{y}. \tag{4}$$

4.1. Hypergraph Construction

A hypergraph $\mathbf{H}$ having size of 4251×6000 elements was formed by concatenating 2000 documents associated to the geo-clusters, 350 topics, z_a, and 1901 vocabulary terms, t_a. The vertex set is defined as $V = D \cup Z_a \cup T_a$. The structure of the hypergraph is summarized in Table 1. For each document d_j associated to a geo-cluster, a hyperedge e_1 is inserted, containing 1 in the jth entry of De_1, 1 for the most related topic to d_j, z_a^* in $Z_a e_1$, and 30 ones for the 30 most descriptive terms t_a for z_a^*, in $T_a e_1$. The weight for this hyperedge is $w(e_1) = P(z_a^*|d_j)$.

To capture the geographical proximity, hyperedges $e_2 \in E_2$ are created. For each d_j corresponding to a specific geo-cluster, one hyperedge e_2 is inserted. It contains 1 to the jth entry, associated to d_j and 1 to the entries corresponding to geo-clusters being at a geographical distance less than 150 km. The weight for this hyperedge is set to 1.

In order to capture the visual similarity of the geo-clusters, the mean value of the GIST descriptors of all the images belonging in a geo-cluster is computed, as a codevector. For each d_j, one e_3 is inserted, having 1 to the jth entry associated to d_j and 1 to the 10 nearest neighbor geo-clusters, identified by applying KNN on the aforementioned codevectors. The hyperedge weight is set to 1.

Table 1. The hypergraph incidence matrix $\mathbf{H}$.

	e_1	e_2	e_3
D	De_1	De_2	De_3
Z_a	$Z_a e_1$	0	0
T_a	$T_a e_1$	0	0

Let d'_j be the geo-cluster where the test image g_{test} belongs to with respect to its geo-tag. The query vector $y \in \mathbb{R}^{|V|}$ is defined as:

$$y(v) = \begin{cases} 1, & \text{if } v = d'_j \\ \Theta(d'_j, v), & \text{otherwise} \end{cases} \tag{5}$$

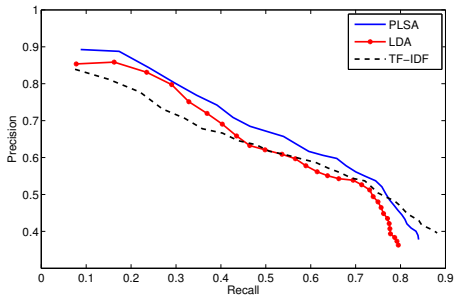

Fig. 3. Recall-precision curves for semantic image annotation by means of PLSA, LDA, and TF-IDF.

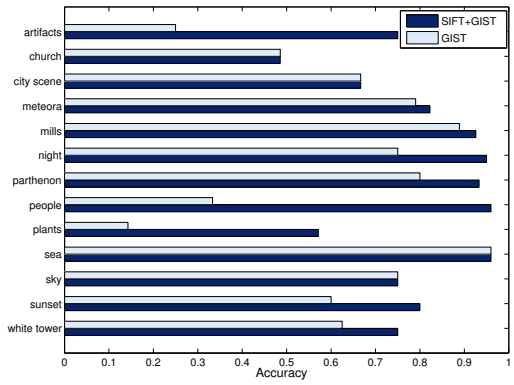

Fig. 4. Accuracy results of the visual image classification.

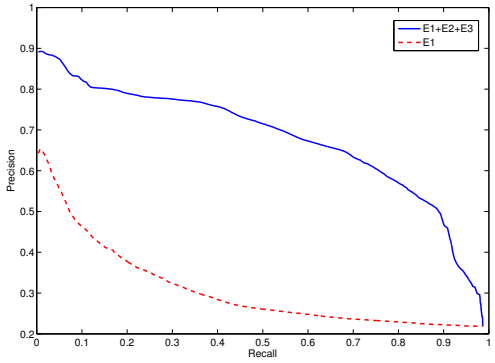

Fig. 5. Recall-precision curves for recommendation.

treating $\Theta(d'_j, v)$ as a measure of relatedness between the vertices of the hypergraph.

The ranking vector $\mathbf{f}^* \in \mathbb{R}^{|V|}$ is derived by solving (4). The values corresponding to the first 2000 entries associated to geo-cluster documents are used as rankings for touristic destination recommendation. The top ranked geo-cluster documents are recommended as touristic POIs to the user, who has imported the test image g_{test}.

5. EXPERIMENTAL RESULTS

For evaluation purposes, a test set containing 205 images was randomly chosen and excluded from the training set along with any text associated to these images. PLSA performance in semantic image annotation has been compared to that of the LDA [10] and the term frequency-inverse document frequency (TF-IDF) [17]. The average recall-precision curve is used as a figure of merit. Precision is defined as the number of correctly recommended objects divided by the number of all recommended objects. Recall is defined as the number of correctly recommended objects, divided by the number of all objects. As it is shown in Fig. 3, PLSA outperforms both LDA and TF-IDF. An average precision of 90% at 10% recall is reported, using PLSA. It is worth noting, that PLSA is much simpler than the LDA.

For visual image classification, the same test set was used. Each test image was assigned into one of 13 representative classes manually in order to form the ground truth. Visual classification accuracy is shown in Fig. 4, when only the GIST descriptors were used and when both SIFT and GIST descriptors were employed, as in Section 3.2. Better results were obtained by using both descriptors. Across the 205 test set images, the average accuracy of content-based image classification over 13 classes is 80%.

Two experiments were conducted to assess touristic POI recommendation. Firstly, only hyperedges $e_1 \in E_1$ were taken into account in hypergraph creation. Secondly, all the hyperedges were considered. The associated recall-precision curves are plotted in Fig. 5. As is clearly indicated, the results are increased when all the three types of hyperedges are considered (including, the visual similarity between the geo-clusters). An average precision of 90% and 82% is reported at 1% and 10% recall, respectively. In order to form the ground truth, relations were established manually among the geo-clusters, taking into account the distance, common geographical entities (e.g., mainland, island) and leisure activities. For this, various tourist related web sources were exploited, such as Trip Advisor[5] and Travel Muse [6].

6. CONCLUSION AND FUTURE WORK

Efficient PLSA driven image annotation, visual image classification, and touristic POI recommendation methods have been proposed and tested on large image collections. The images have been annotated geographically, semantically, and visually by exploiting visual attributes and text information. Furthermore, tourism recommendation has been developed based on hypergraph ranking with promising results. Enhancing these methods by exploiting personalized user information or integrating on-line updating for hyperedge weights could be topics of future research.

Acknowledgments.

This research has been co-financed by the European Union (European Social Fund - ESF) and Greek national funds through the Operation Program "Competitiveness-Cooperation 2011" - Research Funding Program: 11SYN-10-1730-ATLAS.

[5]http://www.tripadvisor.com.gr/
[6]http://www.travelmuse.com/

7. REFERENCES

[1] S. L. Feng, R. Manmatha, and V. Lavrenko, "Multiple Bernoulli relevance models for image and video annotation," in *Proc. IEEE Computer Society Conf. Computer Vision and Pattern Recognition*, 2004, vol. 2, pp. II–1002–II–1009.

[2] C. Wang, D. Blei, and L. Fei-Fei, "Simultaneous image classification and annotation," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2009, pp. 1903–1910.

[3] D. M. Blei and J. D. McAuliffe, "Supervised topic models," in *Proc. Conf. Neural Information Processing Systems*, 2007, vol. 7, pp. 121–128.

[4] X. Zhang, X. Zhao, Z. Li, J. Xia, R. Jain, and W. Chao, "Social image tagging using graph-based reinforcement on multi-type interrelated objects," *Signal Processing*, vol. 93, no. 8, pp. 2178–2189, 2013.

[5] Z. Guan, J. Bu, Q. Mei, C. Chen, and C. Wang, "Personalized tag recommendation using graph-based ranking on multi-type interrelated objects," in *Proc. 32nd Int. ACM SIGIR Conf. Research and Development in Information Retrieval*, 2009, pp. 540–547.

[6] L. Cao, J. Luo, A. Gallagher, X. Jin, J. Han, and T. S. Huang, "A worldwide tourism recommendation system based on geotagged web photos," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, 2010, pp. 2274–2277.

[7] T. Hofmann, "Probabilistic latent semantic indexing," in *Proc. 22nd Int. ACM SIGIR Conf. Research and Development in Information Retrieval*, 1999, pp. 50–57.

[8] N. Bassiou and C. Kotropoulos, "RPLSA: A novel updating scheme for probabilistic latent semantic analysis," *Computer Speech & Language*, vol. 25, no. 4, pp. 741–760, 2011.

[9] N. Bassiou and C. Kotropoulos, "On-line PLSA: Batch updating techniques including out of vocabulary words," *IEEE Trans. Neural Networks and Learning Systems*, 2014, to appear.

[10] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[11] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. Journal Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[12] A. Oliva and A. Torralba, "Building the GIST of a scene: The role of global image features in recognition," *Progress in Brain Research*, vol. 155, pp. 23–36, 2006.

[13] C. Berge and E. Minieka, *Graphs and Hypergraphs*, vol. 7, North-Holland, Amsterdam, 1973.

[14] A. Bosch, A. Zisserman, and X. Munoz, "Scene classification via PLSA," in *Proc. European Conf. Computer Vision*, pp. 517–530. 2006.

[15] S. Agarwal, K. Branson, and S. Belongie, "Higher order learning with graphs," in *Proc. 23rd Int. Conf. Machine Learning*, 2006, pp. 17–24.

[16] J. Bu, S. Tan, C. Chen, C. Wang, H. Wu, Z. Lijun, and X. He, "Music recommendation by unified hypergraph: Combining social media information and music content," in *Proc. ACM Conf. Multimedia*, 2010, pp. 391–400.

[17] G. Salton, A. Wong, and C. Yang, "A vector space model for automatic indexing," *Communications of the ACM*, vol. 18, no. 11, pp. 613–620, 1975.