Tripartite Graph Models for Multi Modal Image Retrieval

Chandrika Pulla
chandrika@research.iiit.ac.in
C.V.Jawahar
jawahar@iiit.ac.in

Center for Visual Information Technology,
IIIT-Hyderabad, INDIA
http://cvit.iiit.ac.in

Most of the traditional image retrieval methods use either low level visual features or embedded text for representation and indexing. In recent years, there has been significant interest in combining these two different modalities for effective retrieval. In this paper, we propose a tri-partite graph based representation of the multi model data for image retrieval tasks. Our representation is ideally suited for dynamically changing or evolving data sets, where repeated semantic indexing is practically impossible. An undirected tripartite graph $G = (T,V,D,E)$ has three sets of vertices where, $T = \{t_1,t_2\ldots,t_n\}$ are text words, $V = \{v_1,v_2\ldots,v_m\}$ are visual words and $D = \{d_1,d_2\ldots,d_i\}$ are images with $E = \{e^{d_1}_{t_1},..,e^{d_i}_{t_n},e^{d_1}_{v_1},..,e^{d_i}_{v_m},e^{t_1}_{v_1},..,e^{t_n}_{v_m}\}$ as set of edges. Figure 1 pictorially represent the tripartite graph model (TGM) we use.

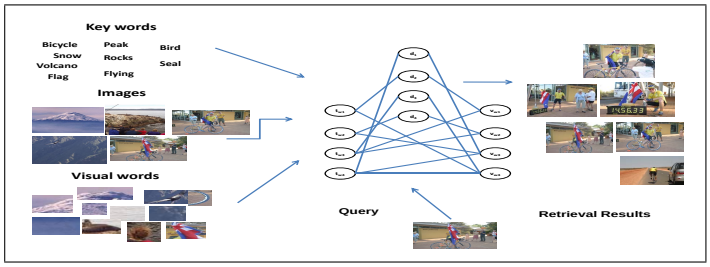

Figure 1: Tri-partite Graph Representation of data set, t_{w_i} are text words, v_{w_i} are visual words and d_i are the images

Thus this model has three sets of vertices (images, text words and visual words) and edges going from one set to other. The nodes correspond to visual words as well as text words store the inverse document frequency (IDF) corresponding to the document(image) collection. The edges from text words to images as well as those from visual words to images, encode the term frequency (TF) corresponding to the word-image pair. However, the weights of edges which relate the text words with visual words can not be directly assigned. These edges are weighed as:

$$W_{pq} = \frac{\sum_i C_{t_p,v_p}(\alpha e^{d_i}_{t_p} + (1-\alpha)e^{d_i}_{v_q})}{\sum_i \alpha e^{d_i}_{t_p} + (1-\alpha)e^{d_i}_{v_q}}$$

Where $C_{t_p,v_q} = 1$, if t_p and v_q are there in document d_i. Since the documents (images) are the entity which connects text words and visual words, summations are carried out over the images/documents.

For indexing, a tripartite graph G is constructed with the nodes and edges as mentioned above. Given a collection of images and textual tags, building a TGM is possible. However, when additional images come, TGM shows its advantage in insertion. To insert an additional image, the TF and IDFs are computed with the new document. We assume the vocabularies to be static. This step is computationally light. For retrieval, we partition the vertex set D of G into two vertex sets $(V1,V2)$, such that vertex set $V1$ contains documents which are relevant to the query, and $V2$ contains all other nodes. For this we employ a graph partitioning procedure explained in Algorithm 1 for retrieving semantically relevant images from the database of images. Being "just in time semantic indexing", our method is computationally light and less resource intensive.

For experimental evaluation, we implemented other Multi modal methods Multi modal LSI(MMLSI), Multi Modal pLSA(MMpLSA) and a multi-layer multi modal LSA as explained in [1, 5]. At first, we construct a Tripartite graph where the edge weights are determined by the widely used Term Frequency. Though it is simple the quality of the similarity measure is not domain dependent and cannot be easily adjusted to better fit the final objective. Therefore we used the method proposed by [7] to learn the edge weights of the tripartite graph to improve the retrieval performance. The Figure 2 shows the comparison of these methods in mean Average Precision(mAP) values. We used four data sets, University

Algorithm 1 Graph Partitioning Algorithm for Tripartite Graph

def **GP**(G,N,R)
Update amount of Relevance score R that have passed through node N
R$[N]$ += R
if Node N is of type Word **then**
  $R = R$ * IDF(N_d) is propagated to document.
  R=R * e^t_v is propagated to visual nodes or text nodes.
end if
if Amount of labels transferable from N $<$ cutoff **then**
  exit
end if
for each *node* **in** neighbourhood of N **do**
  GP$(G,node,R*TF(N,node))$
end for

of Washington(UW) Data set [4], Multi-label Image Data set [6], IAPR TC12 Data set [3], NUS-WIDE [2] for the evaluation of the methods proposed. For all our experiments the number of concepts is determined by the concepts present in the respective databases that are known. The mAP results show that performance of TGM is comparable to other methods. The performance of TGM with weighted-learning is slightly better that that with the TF. The advantage of TGM is noticeable when new images are added to database. TGM takes only few milliseconds for semantic indexing whereas for variants of pLSA the entire semantic indexing needs to be done again, incurring high time and memory costs.

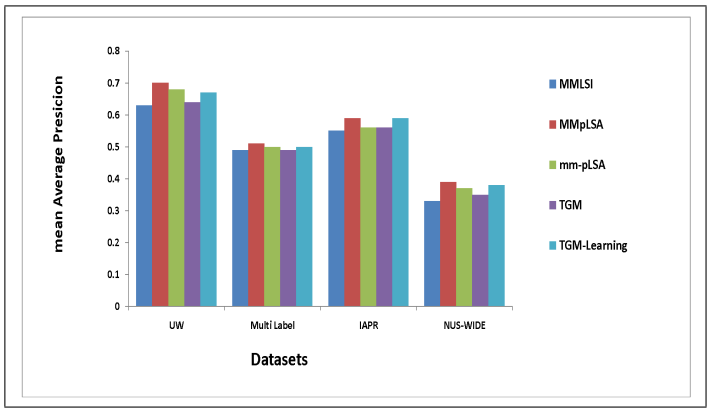

Figure 2: Comparing TGM with Multi Modal LSI and Multi Modal pLSA for different the data sets

[1] Pulla Chandrika and C. V. Jawahar. Multi modal semantic indexing for image retrieval. In *CIVR*, 2010.

[2] Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yan-Tao.Zheng. Nus-wide: A real-world web image database from national university of singapore. In *CIVR*, 2009.

[3] M. Grubinger. Analysis and evaluation of visual information systems performance. In *PhD thesis, Victoria University Melbourne, Australia*, 2007.

[4] Changhu Wang. Lei Zhang, Hong-Jiang Zhang. Scalable markov model-based image annotation. In *IJCV*, 2004.

[5] Rainer Lienhart, Stefan Romberg, and Eva Hörster. Multilayer plsa for multimodal image retrieval. In *CIVR*, 2009.

[6] Singh, P. M., Cunningham, and E. Curran. Active learning for multi-label image annotation. In *AICS*, 2008.

[7] Wen-tau Yih. Learning term-weighting functions for similarity measures. In *EMNLP*, 2009.