Online Multimodal Co-indexing and Retrieval of Weakly Labeled Web Image Collections

Lei Meng[1], Ah-Hwee Tan[2], Cyril Leung[1,3], Liqiang Nie[4],
Tat-Seng Chua[4], Chunyan Miao[1,2,*]
[1]Joint NTU-UBC Research Centre of Excellence in Active Living for the Elderly (LILY), Nanyang Technological University, Singapore
[2]School of Computer Engineering, Nanyang Technological University, Singapore
[3]Department of Electrical and Computer Engineering, The University of British Columbia, Canada
[4]School of Computing, National University of Singapore, Singapore
{lmeng,asahtan,ascymiao,cleung}@ntu.edu.sg, nieliqiang@gmail.com, dcscts@nus.edu.sg

ABSTRACT

Weak supervisory information of web images, such as captions, tags, and descriptions, make it possible to better understand images at the semantic level. In this paper, we propose a novel online multimodal co-indexing algorithm based on Adaptive Resonance Theory, named OMC-ART, for the automatic co-indexing and retrieval of images using their multimodal information. Compared with existing studies, OMC-ART has several distinct characteristics. First, OMC-ART is able to perform online learning of sequential data. Second, OMC-ART builds a two-layer indexing structure, in which the first layer co-indexes the images by the key visual and textual features based on the generalized distributions of clusters they belong to; while in the second layer, images are co-indexed by their own feature distributions. Third, OMC-ART enables flexible multimodal search by using either visual features, keywords, or a combination of both. Fourth, OMC-ART employs a ranking algorithm that does not need to go through the whole indexing system when only a limited number of images need to be retrieved. Experiments on two published data sets demonstrate the efficiency and effectiveness of our proposed approach.

Categories and Subject Descriptors

H.3.1 [**Content Analysis and Indexing**]: Indexing methods; H.3.3 [**Information Search and Retrieval**]: Clustering, Information filtering.

General Terms

Algorithms, Theory, Experimentations

Keywords

Hierarchical image co-indexing, multimodal search, online learning, clustering, weakly supervised learning.

*Corresponding author.

1. INTRODUCTION

Automatic indexing and retrieval of images based on visual features has been a widely studied problem. However, due to the diverse visual content of images, the low-level visual features are typically not consistent in modeling the characteristics of images belonging to the same class, a problem known as the semantic gap [19, 2]. Recently, a number of studies make use of the surrounding text of images, such as captions, tags, and descriptions, as additional features for better understanding and representation of images [21, 15, 22, 7, 11, 20, 10, 18, 17]. The surrounding text of images, also referred to as weak supervision [8], side information [16], and meta-information [15], typically involves high-level semantics that describe the background, objects, and even events about the images. On one hand, using multimodal information helps to improve the indexing and retrieval performance of images. On the other hand, multimodal indexing facilitates multimodal image search based on visual content, keywords, or their combination [9, 1].

Existing studies on multimodal image indexing and retrieval typically focus on techniques that can either identify a latent feature space for the image representations by fusing the multimodal feature representations, such as the Latent Semantic Indexing (LSI) [2, 3], probabilistic Latent Semantic Analysis (pLSA) [10, 3], and Non-negative Matrix Factorization (NMF) [1], or infer the associations among the multimodal features in order to generate a new representation for each image [7, 20, 9]. However, several limitations of such approaches have been identified. First, these approaches cannot perform online learning. Therefore, they cannot handle the large live streams of images that require frequent updates. Second, the surrounding text of images typically has several descriptive keywords together with a relatively large number of words that are not descriptive to the image content [15]. Such noisy information may result in spurious relation between images and have side-effect on the distribution of images in the derived feature space. Third, some of the existing approaches support only one type of queries while those supporting multiple types of queries typically require the generation of multiple transformation matrices, which limits their flexibility for multimodal search. Lastly, all existing approaches return the search result to the query by directly ranking the images of the whole data set. This results in a slow response time for a given query when the data set is large.

To address the aforementioned issues, we propose an online unsupervised learning algorithm, named Online Multimodal Co-indexing Adaptive Resonance Theory (OMC-ART), for the automatic multimodal co-indexing and retrieval of weakly labeled web image collections. In contrast to existing approaches, OMC-ART performs online learning, which allows the adaptation of the learnt indexing system, rather than a re-indexing of the whole data set that will incur heavy computation. To alleviate the side-effect of noisy information and reduce computation complexity, OMC-ART formulates the indexing process as that of simultaneously identifying the clusters of similar images and the key features from their generalized feature distributions, in terms of visual and textual features. As such, OMC-ART generates a two-layer indexing structure, wherein images are co-indexed by the key visual and textual features based on the generalized distributions of the clusters in the cluster-level layer, named the abstraction layer; and their own feature distribution in the image-level layer, named the object layer. Moreover, OMC-ART enables multimodal search by using either visual features, keywords, or a combination of both; and employs a ranking algorithm that iteratively selects the most similar cluster in the abstraction layer and subsequently sorts the images therein in a ranked list. This ranking algorithm may reduce the computational cost because of the pre-ranking of clusters, and will be more efficient when only a limited number of images need to be retrieved.

We evaluate the performance of OMC-ART using two published web image datasets, namely, the NUS-WIDE and Corel5k data sets. In the experiments, we report our studies on parameter selection, retrieval performance comparison, and efficiency analysis. The experimental results show that OMC-ART has a much better performance in terms of the mean Average Precision, Precision, and Recall, and has a much faster response time.

The remainder of this paper is organized as follows. Section 2 summarizes existing studies on multimodal image indexing and retrieval. Section 3 presents the problem formulation. The technical details and experimental evaluation of OMC-ART are described in Section 4 and Section 5, respectively. Section 6 summarizes the main findings of the study and suggests several extensions.

2. RELATED WORK

Multimodal image indexing and retrieval typically follow two main approaches. The first approach is to extend existing algorithms for image indexing with single type of features for integrating multiple types of features. Examples include Latent Semantic Indexing (LSI) [2, 3], probabilistic Latent Semantic Analysis (pLSA) [10, 3], and Non-negative Matrix Factorization (NMF) [1]. Caicedo et al. [2] proposed a Latent Semantic Kernel (LSK), based on LSI, which adopts kernel methods to compute the similarity between the query and the indexed images. Multimodal LSI (MMLSI) [3] utilizes tensors for multimodal image representation and employs Higher Order Singular Value Decomposition (HOSVD) [5] for obtaining the feature representation of images. Chandrika et al. [3] extended pLSA by jointly considering visual and textual features in a probabilistic model, and employed EM algorithm to obtain the derived representation of the images. The Multilayer Multimodal probabilistic Latent Semantic Analysis (MM-pLSA) [10] handles the visual and textual information of images by a multi-layer model, which consists of two leaf pLSA models for learning the visual and textual representation of images respectively, and a node pLSA for obtaining a unified representation. Caicedo et al. [1] proposed two methods based on Non-negative Matrix Factorization (NMF), of which the first method concatenates the matrices for visual and textual features in order to enable search by both visual and textual features, while the second method aims to successively optimize the transformation matrices of textual and visual features, which enables search by using either visual features or keywords.

The second approach is to construct a new representation by exploring the association among multimodal features. Li et al. [9] proposed four methods to infer the similarity matrices for the visual and textual features. The learned similarities are utilized for tackling image retrieval based on visual or textual features. Escalante et al. [7] proposed two methods for image indexing based on the occurrences and co-occurrences information of terms in the surrounding text and the object labels associated to images. The hybrid framework [20], named iSMIER, performs image retrieval by predicting the captions and annotations for the query image, and indexing it by its visual fuzzy membership of clusters.

3. PROBLEM FORMULATION

Given a weakly labeled image collection $\mathcal{D} = \{d_1, ..., d_N\}$ with their associated visual features $\mathcal{V} = \{\mathcal{V}_1, ..., \mathcal{V}_M\}$ and textual features $\mathcal{T} = \{\mathcal{T}_1, ..., \mathcal{T}_H\}$, the nth image is represented by the visual feature vector $\mathbf{v}_n = [v_{n,1}, ..., v_{n,M}]$, and the textual feature vector $\mathbf{t}_n = [t_{n,1}, ..., t_{n,H}]$ $(n = 1, ..., N)$.

The multimodal co-indexing and retrieval problem of weakly labeled images is defined as the process of simultaneously identifying a set of clusters of similar images $\mathcal{C} = \{c_1, ..., c_J\}$ and their generalized visual and textual feature representations as weight vectors $\mathbf{w}_j^v = [w_{j,1}^v, ..., w_{j,M}^v]$ and $\mathbf{w}_j^t = [w_{j,1}^t, ..., w_{j,H}^t]$ $(j = 1, ..., J)$. In this way, the key visual and textual features of each cluster c_j $(j = 1, ..., J)$ can be identified according to the cluster weight vectors $\mathbf{w}_j^v$ and $\mathbf{w}_j^t$, denoted as $\mathcal{K}_j^v = \{v_m | v_m \in key\,features\,of\,c_j\}$ and $\mathcal{K}_j^t = \{t_m | t_m \in key\,features\,of\,c_j\}$, respectively. As such, images in the indexing system, e.g. the nth image $d_n \in c_j$, will be co-indexed by the identified key visual features $\mathcal{K}_n^v = \{v_{n,m} | v_m \in \mathcal{K}_j^v\}$ and textual features $\mathcal{K}_n^t = \{t_{n,h} | t_h \in \mathcal{K}_j^t\}$.

The subsequent retrieval problem is defined as a ranking process. Specifically, a query q can either be an image, or several keywords, or a combination of both. When q is presented to OMC-ART, the corresponding visual and textual feature vectors $\mathbf{v}_q$ and $\mathbf{t}_q$ will be constructed based on $\mathcal{V}$ and $\mathcal{T}$. By calculating the similarities between the query and images in the indexing system $S(q, d_n)$, a list of most similar images $\mathcal{L}$ is returned as the retrieval result.

4. ONLINE MULTIMODAL CO-INDEXING ADAPTIVE RESONANCE THEORY

Online Multimodal Co-indexing Adaptive Resonance Theory (OMC-ART) comprises three steps. First, OMC-ART employs an adaptation method to extend the heterogeneous data co-clustering algorithm, named Generalized Heterogeneous Fusion Adaptive Resonance Theory (GHF-ART) [15], to perform online learning and generate the clusters of similar images with the respective generalized visual and textual feature distributions. Second, making use of the learnt weight vectors of the discovered image clusters, OMC-

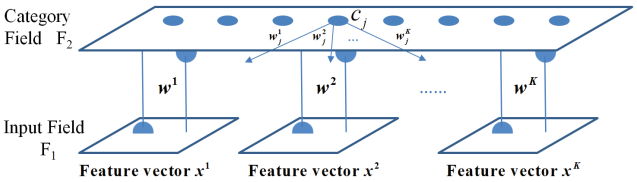


Figure 1: The architecture of GHF-ART for integrating K types of feature vectors.

ART dynamically selects the key features of each cluster in order to co-index the images in the clusters using a two-layer hierarchical indexing structure. Third, OMC-ART incorporates a ranking algorithm that allows multiple types of queries for retrieving images in the indexing system.

4.1 GHF-ART

4.1.1 Heterogeneous Feature Representation

GHF-ART [15], as shown in Figure 1, consists of K independent feature channels in the input field, each of which handles one feature modality $\mathbf{x}^k$ of the input data object, and a category field consisting of clusters that are represented by weight vectors $\mathbf{w}^k$ $(k = 1, ..., K)$. It allows for different representation and learning methods for different feature modalities. Regarding the weakly labeled image collection, given an image d_n, GHF-ART may receive any type of visual features in the form of a vector $\mathbf{v}_n = [v_{n,1}, ..., v_{n,M}]$, which should be further normalized by the *min-max normalization* that guarantees the input values are in the interval $[0, 1]$. The corresponding textual feature vector $\mathbf{t}_n = [t_{n,1}, ..., t_{n,H}]$ is represented by the presence of words in d_n, defined by

$$t_{n,h} = \begin{cases} 1, & if\ t_h \in d_n \\ 0, & otherwise \end{cases}. \tag{1}$$

4.1.2 Clustering Procedures of GHF-ART

GHF-ART performs clustering of composite data objects in an incremental manner. Given an input data object d_n represented by K types of features $\mathcal{I}_n = \{\mathbf{x}_n^k|_{k=1}^K\}$, the clustering process of GHF-ART has three steps:

1) Category Choice: When the input vectors $\mathcal{I}_n$ of the data object d_n is presented, a choice function is firstly employed to evaluate the overall similarity between d_n and each cluster c_j in the category field, which is defined by

$$T(d_n, c_j) = \sum_{k=1}^{K} \gamma^k \frac{|\mathbf{x}_n^k \wedge \mathbf{w}_j^k|}{\alpha + |\mathbf{w}_j^k|}, \tag{2}$$

where $\mathbf{w}_j^k$ denotes the weight vector of the jth cluster for the kth feature modality, the contribution parameter $\gamma^k \in [0, 1]$ is the weight for the kth feature modality, the choice parameter $\alpha \approx 0$ is a positive real value to balance the denominator, the operation $\wedge$ is defined by $(\mathbf{x}_n^k \wedge \mathbf{w}_j^k)_i \equiv \min(x_{n,i}^k, w_{j,i}^k)$, and $|.|$ is the ℓ_1 norm.

2) Template Matching: After identifying the winner cluster c_{j^*}, a match function is used to evaluate if c_{j^*} matches d_n in terms of each feature modality. For the kth feature modality, the match function is defined by

$$M(\mathbf{x}_n^k, \mathbf{w}_{j^*}^k) = \frac{|\mathbf{x}_n^k \wedge \mathbf{w}_{j^*}^k|}{|\mathbf{x}_n^k|}. \tag{3}$$

3) New Cluster Creation or Resonance: Given the vigilance parameter $\rho^k \in [0, 1]$ for the kth feature modality, if the vigilance criteria $M(\mathbf{x}_n^k, \mathbf{w}_{j^*}^k) > \rho^k$ for $k = 1, ...K$ are not satisfied, a reset occurs so that a new winner is selected. If all clusters in the category field do not meet the vigilance criteria, a new cluster, denoted as c_{new}, will be created to encode d_n such that $\mathbf{w}_{c_{new}}^k = \mathbf{x}_n^k$ for $k = 1, ..., K$. Otherwise, a resonance occurs so that the weight vectors $\{\mathbf{w}_{j^*}^k|_{k=1}^K\}$ of c_{j^*} are updated in terms of different modalities. Assuming that $\mathbf{w}_{j^*}^k$ contains the features extracted from images, the corresponding update equation is defined by

$$\hat{\mathbf{w}}_{j^*}^k = \beta(\mathbf{x}_n^k \wedge \mathbf{w}_{j^*}^k) + (1 - \beta)\mathbf{w}_{j^*}^k, \tag{4}$$

where $\beta \in [0, 1]$ is the learning parameter. In contrast, assuming that $\mathbf{w}_{j^*}^k$ contains the features extracted from the surrounding text of images, the update equation is defined by

$$\hat{w}_{j^*,h}^k = \begin{cases} \eta w_{j^*,h}^k & if\ x_{n,h}^k = 0 \\ \eta(w_{j^*,h}^k + \frac{1}{L}) & otherwise \end{cases}, \tag{5}$$

where $w_{j^*,h}^k$ is the hth feature of $\mathbf{w}_{j^*}^k$, $x_{n,h}^k$ is the hth feature of $\mathbf{x}_n^k$, L is the number of data objects in c_{j^*}, and $\eta = \frac{L}{L+1}$.

4.1.3 Adaptive Weighting for Heterogeneous Features

GHF-ART adaptively tunes the weights γ^k $(k = 1, ...K)$ for different feature modalities as used in Equation (2) through the *Robustness Measure*, which evaluates the importance of different feature modalities in recognizing similar data objects. Considering a cluster c_j with L data objects, each of which is denoted by $\mathcal{I}_l = \{\mathbf{x}_l^1, \ldots, \mathbf{x}_l^K\}$ for $l = 1, \ldots, L$, and the corresponding weight vectors for the K feature modalities are denoted by $\mathcal{W}_j = \{\mathbf{w}_j^1, \ldots, \mathbf{w}_j^K\}$, the *Difference* for the kth feature modality of c_j is measured by

$$D_j^k = \frac{\frac{1}{L}\sum_l |\mathbf{w}_j^k - \mathbf{x}_l^k|}{|\mathbf{w}_j^k|}. \tag{6}$$

Considering all the J clusters, the *Robustness* of the kth feature modality can be measured by

$$R^k = \exp(-\frac{1}{J}\sum\nolimits_j D_j^k). \tag{7}$$

Finally, the contribution parameter for the kth feature channel γ^k is defined by

$$\gamma^k = \frac{R^k}{\sum_{k=1}^K R^k}. \tag{8}$$

For efficiency purpose, the respective incremental update equations for γ^k $(k = 1, ...K)$ are further derived:

- **Resonance in existing cluster**: Assuming the input data object d_{L+1} with feature vectors $\mathcal{I}_{L+1} = \{\mathbf{x}_{L+1}^1, \ldots, \mathbf{x}_{L+1}^K\}$ is assigned to the cluster c_j. For the kth feature modality, the update equations for the features extracted from images and surrounding text are defined by equations (9) and (10) respectively:

$$\hat{D}_j^k = \begin{cases} \frac{\eta}{|\hat{\mathbf{w}}_j^k|}(|\mathbf{w}_j^k|D_j^k + |\mathbf{w}_j^k - \hat{\mathbf{w}}_j^k| + \frac{1}{L}|\hat{\mathbf{w}}_j^k - \mathbf{x}_{L+1}^k|), & (9) \\ \frac{\eta}{|\hat{\mathbf{w}}_j^k|}(|\mathbf{w}_j^k|D_j^k - |\hat{\mathbf{w}}_j^k - \eta\mathbf{w}_j^k| + \frac{1}{L}|\hat{\mathbf{w}}_j^k - \mathbf{x}_{L+1}^k|) & (10) \end{cases}$$

  After the update for all feature modalities, the updated contribution parameter can then be obtained using Equations (7) and (8).

- **Generation of new cluster**: When generating a new cluster, the differences of other clusters remain unchanged. Therefore, it just introduces a proportional change in *Difference*. Considering the *robustness* R^k $(k = 1, ..., K)$ for all feature modalities, the update contribution parameter for the kth feature modality is defined by:

$$\hat{\gamma}^k = \frac{(R^k)^{\frac{J}{J+1}}}{\sum_{k=1}^{K}(R^k)^{\frac{J}{J+1}}}. \quad (11)$$

4.2 Online Adaptation of Normalized Feature Distributions

GHF-ART may not be directly applicable to online learning, because the *min-max normalization* requires the maximum and minimum values of each feature for normalizing the features extracted from document content. To address this issue, OMC-ART employs an adaptation method that updates the normalized feature vectors of data objects and cluster weights to what they should be when an input data object incurs a change in such values, as defined by Equations (12) and (13) below,

$$x^{(new)} = \frac{x_{max}^{(old)} - x_{min}^{(old)}}{x_{max}^{(new)} - x_{min}^{(new)}} x^{(old)} + \frac{x_{min}^{(old)} - x_{min}^{(new)}}{x_{max}^{(new)} - x_{min}^{(new)}}, \quad (12)$$

$$w^{(new)} = \frac{x_{max}^{(old)} - x_{min}^{(old)}}{x_{max}^{(new)} - x_{min}^{(new)}} w^{(old)} + \frac{x_{min}^{(old)} - x_{min}^{(new)}}{x_{max}^{(new)} - x_{min}^{(new)}}, \quad (13)$$

where x denotes a feature extracted from document content, $x^{(old)}$ is the value of x calculated based on the old maximum value $x_{max}^{(old)}$ and minimum value $x_{min}^{(old)}$, and $x^{(new)}$ is the updated value calculated based on the new maximum value $x_{max}^{(new)}$ and minimum value $x_{min}^{(new)}$.

As an online algorithm, the initial maximum and minimum values $x_{max}^{(1)}$ and $x_{min}^{(1)}$ should be carefully considered. Without the loss of generalization, we set $x_{max}^{(1)} = x^{(0)}$ and $x_{min}^{(1)} = x^{(0)} - 1$, where $x^{(0)}$ is the original value of x without normalization.

THEOREM 1. *Considering a feature x that will be continuously normalized by N set of maximum and minimum values $\{x_{max}^{(n)}, x_{min}^{(n)}\}_{n=1}^{N}$, the value of x with n round of normalization $x^{(n)}$ can be inferred directly by that of $x^{(n-1)}$ by Equation (12).*

PROOF. Given $x_{max}^{(n)}$ and $x_{min}^{(n)}$, we have

$$x^{(n)} = \frac{x^{(0)} - x_{min}^{(n)}}{x_{max}^{(n)} - x_{min}^{(n)}}, \quad (14)$$

$$x^{(n-1)} = \frac{x^{(0)} - x_{min}^{(n-1)}}{x_{max}^{(n-1)} - x_{min}^{(n-1)}}. \quad (15)$$

By substituting $x^{(0)}$ in Equation (14) using the expression of $x^{(0)}$ derived from Equation (15), we have

$$x^{(n)} = \frac{x_{max}^{(n-1)} - x_{min}^{(n-1)}}{x_{max}^{(n)} - x_{min}^{(n)}} x^{(n-1)} + \frac{x_{min}^{(n-1)} - x_{min}^{(n)}}{x_{max}^{(n)} - x_{min}^{(n)}}. \quad (16)$$

□

THEOREM 2. *Considering the value of weight w, denoted by $w^{(N)}$, which learns from the features $\{x_n^{(N)}\}_{n=1}^{N}$ of N data objects, and is normalized by $\{x_{max}^{(N)}, x_{min}^{(N)}\}$. If a new input data object d_{N+1} introduces $x_{max}^{(N+1)}$ and $x_{min}^{(N+1)}$, the adapted weight value $w^{(N+1)}$ can be derived by Equation (13).*

PROOF. Section 4.1.2 indicates that $w^{(1)} = x_1^{(N)}$, and the value of w is updated by learning from input data objects following Equation (4). Therefore, we have

$$w^{(n)} = \begin{cases} w^{(n-1)} & if\ x_{n-1}^{(N)} \geq w^{(n-1)} \\ (1-\beta)w^{(n-1)} + \beta x_n^{(N)} & otherwise \end{cases}. \quad (17)$$

Based on Equation (17), we may infer that $w^{(N)} = c_1 x_1^{(N)} + ... + c_N x_N^{(N)}$ and $c_1 + ... + c_N = 1$. Therefore, when $x_{max}^{(N+1)}$ and $x_{min}^{(N+1)}$ are introduced, we obtain $w^{(N+1)} = c_1 x_1^{(N+1)} + ... + c_N x_N^{(N+1)}$. By denoting Equation (16) by $x^{(n)} = a^{(n)} x^{(n-1)} + b^{(n)}$, we have

$$\begin{aligned} w^{(N+1)} &= a^{(N+1)}(c_1 x_1^{(N)} + ... + c_N x_N^{(N)}) + (c_1 + ... + c_N) b^{(N+1)} \\ &= a^{(N+1)} w^{(N)} + b^{(N+1)}. \end{aligned} \quad (18)$$

□

4.3 Dynamic Selection of Key Features and Hierarchical Co-Indexing of Images

The learning functions of GHF-ART for document content and surrounding text, as discussed in Section 4.1.2, essentially aim to discover the key features by preserving or increasing the values of key features while decreasing those of noisy features. Therefore, the representative visual and textual features $\mathcal{V}$ and $\mathcal{T}$ can be obtained from the corresponding cluster weight vectors $\{\mathbf{w}_j^k\}_{j=1}^{J}$ $(k = \{v, t\})$ of the clusters $\{c_j\}_{j=1}^{J}$. The key visual and textual features $\mathcal{K}_j^v$ and $\mathcal{K}_j^v$ of c_j are selected based on the following criteria,

$$\mathcal{K}_j^v = \{v_m | w_{j,m}^v > \frac{1}{M} \sum_{i=1}^{M} w_{j,i}^v\}, \quad (19)$$

$$\mathcal{K}_j^t = \{t_h | w_{j,h}^t > \frac{1}{H} \sum_{i=1}^{H} w_{j,i}^t\}. \quad (20)$$

The proposed criteria select the features of values above average as key features. They are based on the idea that the high dimensional features are usually sparse and noisy, especially for the surrounding text of images. Therefore, the proposed method may filter the features providing little information while keeping those that are useful for indicating the difference between clusters.

In this way, each image $d_n \in c_j$ in the indexing system of OMC-ART is hierarchically indexed by the key weight values, $\mathcal{KW}_j^v = \{w_{j,m}^v | v_m \in \mathcal{K}_j^v\}$ and $\mathcal{KW}_j^t = \{w_{j,h}^t | t_h \in \mathcal{K}_j^t\}$, in the abstraction layer and the corresponding feature values, $\mathcal{K}_n^v = \{v_{n,m} | v_m \in \mathcal{K}_j^v\}$ and $\mathcal{K}_n^t = \{t_{n,h} | t_h \in \mathcal{K}_j^t\}$, in the object layer.

4.4 Ranking for Multimodal Queries

OMC-ART enables multimodal search by using either visual features, keywords, or combination of both. Given a query q, the visual and/or textual feature vectors, $\mathbf{v}_q$ and $\mathbf{t}_q$, for the provided query image and/or keywords will be constructed based on $\mathcal{V}$ and $\mathcal{T}$. Taking advantage of the two-layer

indexing structure, we employ a ranking algorithm based on binary insertion sort. In the first step, the similarity between the query q and the clusters c_j for $j = 1, ..., J$ in the abstraction layer will be computed. We first define the dissimilarity between two feature values as

$$DIS(a_i, b_i) = \frac{\max(a_i, b_i) - \min(a_i, b_i)}{\alpha + a_i}. \quad (21)$$

The dissimilarity evaluates the degree of the difference between a_i and b_i to a_i, and α is defined in Equation (2). Subsequently, the similarity between q and c_j is defined as

$$\begin{aligned} S_a(q, c_j) = \gamma^v \sum_{w^v_{j,i} \in \mathcal{KW}^v_j} \max(0, 1 - DIS(v_{q,i}, w^v_{j,i})) \\ + \gamma^t \sum_{w^t_{j,i} \in \mathcal{KW}^t_j} \max(0, 1 - DIS(t_{q,i}, w^t_{j,i})), \end{aligned} \quad (22)$$

where γ^v and γ^t are the weights learnt by Equation (8) during clustering, which assign higher weights to the more important feature modality. For queries using either image or keywords, the corresponding part in Equation (22) will not be considered. Here, the max(.) function is utilized to avoid the case that the selected key features of clusters are not the key features of the query.

Given the cluster $c_j \in \mathcal{L}_c$ that is most similar to query q, each $d_n \in c_j$ is inserted to the ranking list $\mathcal{L}$ according to the binary insertion sort. Considering an image $d_n \in c_j$, the similarity between the query q and d_n is defined as

$$\begin{aligned} S_o(q, d_n) = \gamma^v \sum_{v_{n,i} \in \mathcal{K}^v_n} w^v_{j,i} \ \max(0, 1 - DIS(v_{q,i}, v_{n,i})) \\ + \gamma^t \sum_{t_{n,i} \in \mathcal{K}^t_n} w^t_{j,i} \max(0, 1 - DIS(t_{q,i}, t_{n,i})). \end{aligned} \quad (23)$$

Note that the weights for similarities are introduced here to enhance the impact of key features. In addition, with a predefined length u of $\mathcal{L}$, the ranking algorithm may stop without traversing the entire indexing system if the ranking list keeps unchanged for a certain period of time, in view that the images most similar to the query are presented prior to those of lower similarity.

4.5 Computational Complexity Analysis

OMC-ART includes a co-indexing module and a retrieval module. Regarding the co-indexing module, for each data object, OMC-ART first normalizes the features, which requires a time complexity of $O(n_i n_f)$, where n_i denotes the number of images and n_f denotes the total number of features. A change in the bound values of features x_{max} and x_{min} will incur a computation cost of $O(n_i n_f)$ in the worst case. Second, the clustering process of OMC-ART, as demonstrated in [15], has an overall time complexity of $O(n_i n_c n_f)$, where n_c is the number of clusters. Finally, the hierarchical co-indexing process has a time complexity of $O((n_i + n_c)n_f)$. Therefore, the co-indexing module of OMC-ART has a total time complexity of $O(n_i n_c n_f)$.

The retrieval module of OMC-ART includes the construction of features, the similarity evaluation between the query and the indexed images, and the ranking algorithm. The feature construction for the query occurs in real-time. If the ranking list $\mathcal{L}$ includes all images in the data set, the overall time complexity for the similarity measure and ranking is of $O(n_i n_{kf} + n_i log n_i)$, where $n_{kf} < n_f$ is the total number of key features. In our experiment on the NUS-WIDE data set, we have $n_f = 2000$ while n_{kf} of a cluster is typically less than 30. In contrast, if $\mathcal{L}$ has a limited length u, the overall time complexity is $O(n_u n_{kf} + n_u log n_u)$, where $n_u < n_i$ is the number of images used to achieve a stable $\mathcal{L}$.

Algorithm 1 OMC-ART - Co-Indexing

Input: Images $\{d_n\}_{n=1}^N$ with the corresponding visual and textual features $\{\mathbf{v}_n\}_{n=1}^N$ and $\{\mathbf{t}_n\}_{n=1}^N$, and parameters $\alpha = 0.001$, $\beta = 0.6$, ρ^v and ρ^t.

Present d_1 with $\mathbf{v}_1$ and $\mathbf{t}_1$ to the input field.
Initialize x_{max} and x_{min} for each feature x of $\mathbf{v}_1$, and perform min-max normalization on $\mathbf{v}_1$.
Set $J = 1$. Create cluster c_J with $\mathbf{w}^v_J = \mathbf{v}_1$ and $\mathbf{w}^t_J = \mathbf{t}_1$.
Set $n = 2$.
repeat
  Present d_n to the input field.
  If x_{max} and x_{min} are changed, update normalized features according to Equations (12) and (13). Normalize $\mathbf{v}_n$ with x_{max} and x_{min}.
  For $\forall c_j$ $(j = 1, ..., J)$, calculate the choice value $T(d_n, c_j)$ according to Equation (2).
  repeat
    Identify a winner cluster c_{j^*} so that $j^* = \arg\max_{j: c_j \in F_2} T(d_n, c_j)$.
    Calculate the match values $M(\mathbf{v}_n, \mathbf{w}^v_{j^*})$ and $M(\mathbf{t}_n, \mathbf{w}^t_{j^*})$ according to Equation (3).
    If $M(\mathbf{v}_n, \mathbf{w}^k_{j^*}) < \rho^k (k = \{v, t\})$, set $T(d_n, c_{j^*}) = -1$.
  until Identify c_{j^*} such that $M(\mathbf{v}_n, \mathbf{w}^k_{j^*}) > \rho^k$ for $k = \{v, t\}$, or $T(d_n, c_{j^*}) = -1$.
  If $T(d_n, c_{j^*}) \neq -1$, set $d_n \in c_{j^*}$, update $\mathbf{w}^k_{j^*}$ for $k = \{v, t\}$ according to Equations (4) and (5) respectively, and update γ^v and γ^t according to Equations (7)-(10).
  If $T(d_n, c_{j^*}) = -1$, set $J = J + 1$, create a new node c_J such that $\mathbf{w}^v_J = \mathbf{v}_n$ and $\mathbf{w}^t_J = \mathbf{t}_n$, and update γ^v and γ^t according to Equation (11).
  Set $n = n + 1$.
until All images are presented.
Identify key features of clusters according to Equations (19) and (20), and obtain the indexes $\mathcal{KW}^v_j$ and $\mathcal{KW}^t_j$ of clusters $\{c_j\}_{j=1}^J$ and the indexes $\mathcal{K}^v_n$ and $\mathcal{K}^t_n$ of images $\{d_n\}_{n=1}^N$ as discussed in Section 4.3.

Output: Clusters $\{c_j\}_{j=1}^J$, cluster assignment of images $\{A_n\}_{n=1}^N$, and indexes $\mathcal{KW}^k_j$ and $\mathcal{K}^k_n$ for $k = \{v, t\}$.

5. EXPERIMENTS

5.1 Data Sets

We conducted experiments on two data sets. The first is the NUS-WIDE data set [4], consisting of 269,648 images with surrounding text and their ground-truth labels of 81 concepts. We used 16,000 images belonging to 10 classes, including dog, birds, flower, lake, sunset, beach, bridge, cars, coral, and garden, each of which includes 1,600 images. We used a concatenation of Grid Color Moment (255 features), Edge Direction Histogram (73 features) and Wavelet Texture (128 features) as visual features. For the textual

Algorithm 2 OMC-ART - Retrieval

Input: Query q (image, keywords, or combination of both).

1: Construct the visual feature vector $\mathbf{v}_q$ and/or textual feature vector $\mathbf{t}_q$, and present them to the input field.
2: Perform min-max normalization on $\mathbf{v}_q$ based on the current x_{max} and x_{min}. If $\exists i$ such that $v_{q,i} > x_{max}$ or $v_{q,i} < x_{min}$, set $v_{q,i} = 1$ or 0, respectively.
3: Calculate $S_a(q, c_j)$ for $j = 1, ...J$ according to Equation (22), and obtain the ranking list $\mathcal{L}_c = \{c_i\}_{i=1}^{J}$.
4: Set $i = 1$.
5: **repeat**
6: Select cluster $c_i \in \mathcal{L}_c$.
7: **repeat**
8: Select an image $d_n \in c_i$, and calculate $S_o(q, d_n)$ according to Equation (23).
9: Find its ranking in the retrieval list $\mathcal{L}$ using the binary search algorithm.
10: **until** all images $d_n \in c_i$ are presented to $\mathcal{L}$, or $\mathcal{L}$ of length u remains unchanged for a period of time.
11: Set $i = i + 1$.
12: **until** All images are presented, or $\mathcal{L}$ of length u remains unchanged for a period of time.

Output: The list $\mathcal{L}$ of ranked images to the query q.

features, we filtered the surrounding text of images and considered all distinctive and high frequency tags. In total, we selected 2,000 tags and ensured that each selected image is associated with at least five tags. For the retrieval performance evaluation, we used 1,500 images of each class, i.e. 15,000 images, for building the indexing system and the remaining 1,000 images for queries.

The second data set used is the Corel5k data set, which is originally used in [6] for visual object recognition. This data set consists of 5,000 images from 50 Corel Stock Photo CDs, each of which contains 100 images of the same class. Each image is typically manually annotated by 3-4 tags from a dictionary of 374 words. However, our obtained data set contains only 4,999 images with one missing. Similar to the NUS-WIDE data set, we utilized the same visual features and 374 words to build the visual and textual feature vectors. We treated images in the same folder as belonging to the same class, and 10 images of each class were selected as queries while the remaining 4,499 images were used to build the indexing system.

5.2 Evaluation Measures

We evaluated the retrieval performance of OMC-ART using four measures, including the mean Average Precision@k ($mAP@K$), Precision@K, Recall@K, and the response time. Given the number of queries, N, the number of relevant images in the indexing system, M_n, to the nth query, and the number of retrieved images K, $mAP@K$ is defined as

$$mAP@K = \frac{1}{N}\sum_{n=1}^{N} AP_n@K, \tag{24}$$

$$AP_n@K = \frac{1}{\min(M_n, K)}\sum_{k=1}^{K} r_{n,k}\sum_{i=1}^{k}\frac{r_{n,i}}{k}. \tag{25}$$

Note that $AP_n@K$ is the Average Precision (AP) obtained by the nth query, $r_{n,k} = 1$ if the kth retrieved image is relevant to the nth query, and $r_{n,k} = 0$ otherwise. $AP_n@K$ considers positions of the relevant images in the ranking list, and the top ranked relevant images will result in a high performance. It essentially calculates the weighted sum of $Precision_n@k$ for $k = 1, ..., K$ when $K < M_n$, and calculates the weighted sum of $Recall_n@k$ when $K > M_n$. In addition, $Precision_n@K = \sum_{i=1}^{K}\frac{r_{n,i}}{K}$, and $Recall_n@K = \sum_{i=1}^{K}\frac{r_{n,i}}{M_n}$. $Precision@K$ and $Recall@K$ are the respective mean values over N queries.

5.3 Parameter Selection

OMC-ART requires three parameters for GHF-ART in order to build the co-indexing module, namely, the choice parameter α, the learning rate β and the vigilance parameters ρ^v and ρ^t. As demonstrated in several studies [12, 13, 14], the performance of GHF-ART is generally robust to the values of α and β, and $\alpha = 0.01$ and $\beta = 0.6$ are commonly used. Therefore, we consistently used $\alpha = 0.01$ and $\beta = 0.6$ across our experiments on the two data sets.

The vigilance parameter ρ essentially constrains the minimum intra-cluster similarity. As specified in [13], a suitable value of ρ typically results in the generation of a few small clusters, typically 10% of the total number of the generated clusters. Besides, a small cluster typically contains several or tens of data objects. Therefore, the moderate values of ρ^v and ρ^t can be obtained based on the clustering results produced by the respective visual and textual features. Note that ρ^v and ρ^t affect the retrieval performance of OMC-ART in terms of the selection of key features and the accuracy of grouping similar data objects. Therefore, relatively higher values of ρ^v and ρ^t are preferred in order to enhance the accuracy of co-indexing with some increase in the computational cost for building the indexing system. In our experiments, we consistently used $\rho^v = 0.8$ and $\rho^t = 0.3$ for the two data sets.

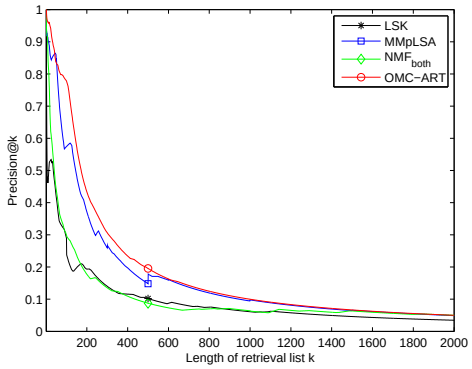
5.4 Performance Comparison

The performance of OMC-ART is compared with four state-of-the-art multimodal images indexing and retrieval algorithms, namely (1) the Latent Semantic Kernels (LSK) [2] which supports query by image, keywords and combination of both; (2) Content-based Image Retrieval ($CBIR$) and Text-based Image Retrieval ($TBIR$) [9] which support query by image and keywords respectively; (3) Multimodal Probabilistic Latent Semantic Analysis ($MMpLSA$) [3] which supports query by combination of image and keywords; and (4) the algorithms based on Non-negative Matrix Factorization (NMF) [1] which support query by image, keywords, and combination of both, denoted as NMF_v, NMF_t, and NMF_{both}, respectively.

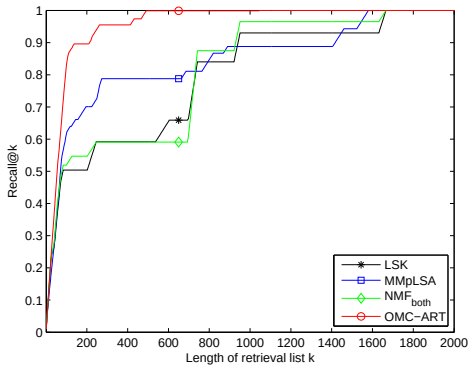
For a fair comparison, we normalized the visual features so that they fit the input of all algorithms. Regarding the algorithm implementations that are not mentioned or have alternatives in the respective papers, for LSK, we used the cosine kernels for feature similarity measure and the linear kernel for combining visual and textual similarities; For CBIR and TBIR, the cosine similarity was used for similarity measure. Because the ranking algorithms of all four algorithms were not mentioned in the respective papers, we used the binary insertion sort as used in OMC-ART. Regarding the parameters such as the weights of features, the number of iterations, and the number of clusters/dimensionality of latent space, we first followed the suggestions in the respec-

Table 1: The retrieval performance of OMC-ART and the baselines on the NUS-WIDE and Corel5k data sets.

mAP / dataset	Query by Image				Query by Keywords				Query by Both			
	LSK	CBIR	NMF_v	OMC-ART	LSK	TBIR	NMF_t	OMC-ART	LSK	MMpLSA	NMF_{both}	OMC-ART
NUS-WIDE	0.1382	0.1763	0.2287	**0.2729**	0.2794	0.3345	0.2936	**0.3804**	0.3474	0.3948	0.3469	**0.4974**
Corel5k	0.1418	0.1976	0.1712	**0.2877**	0.3391	0.3412	0.3682	**0.4865**	0.3552	0.3991	0.3875	**0.5283**



(a)



(b)

Figure 2: The retrieval performance of OMC-ART and the compared algorithms on Corel5k data set with queries of both image and keywords, in terms of (a) Precision@k and (b) Recall@k.

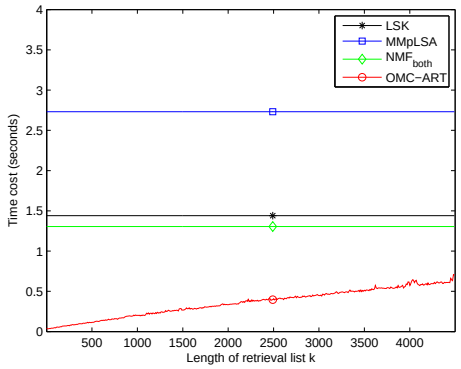


Figure 3: The time cost of OMC-ART and the compared algorithms on Corel5k data set with respect to the increase in the length of retrieval list k.

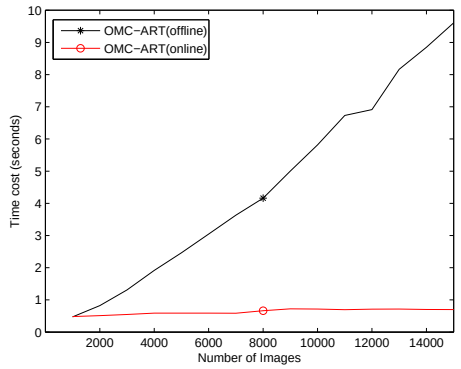


Figure 4: The time cost of OMC-ART with online learning (OMC-ART(online)) and offline learning (OMC-ART(offline)) to build the indexing system with live streams of images from NUS-WIDE data set.

tive papers, and then empirically tuned them so that each algorithm achieved roughly the best retrieval performance in terms of mAP.

Table 1 summarizes the retrieval performances of OMC-ART and the compared algorithms on the NUS-WIDE and Corel5k data sets, evaluated by mAP. We observed that OMC-ART consistently achieved the best performance in terms of all types of queries and data sets, which was usually over 10% higher than that achieved by the compared algorithms. Besides, we found that, when querying with combined image and keywords, the performance of OMC-ART was significantly better than that of querying by using either image or keywords. The above findings demonstrated the effectiveness of the proposed co-indexing method, which indexes the images using the discovered key features of each modality to enhance the accuracy of similarity measure.

In addition, we evaluated the retrieval performance of OMC-ART and the compared algorithms using query by combination of both image and keywords on the Corel5k data set. The performance was measured by Precision@k and Recall@k with respect to the increase in the length of the retrieval list k, as shown in Figure 2. In Figure 2(a), we observed that OMC-ART always obtained the best results of precision with different numbers of k, and the performance decreased slower than other algorithms along with the increase in k. Regarding the recall shown in Figure 2(b), we observed that OMC-ART had a much better recall than other algorithms, and typically identified all images similar to a given query at $k \geq 500$.

5.5 Efficiency Analysis

To demonstrate the efficiency of OMC-ART, we first evaluated the response time of OMC-ART and the compared algorithms on the Corel5k data set using both image and keywords as queries, with respect to the length of the retrieval list k. To make a fair comparison, we empirically tuned the dimensionality of latent space of LSK, $MMpLSA$ and NMF_{both} to be the same under their respective best settings as used in Section 5.4. As illustrated in Figure 3, OMC-ART requires the least time cost among all algorithms, which can be shorter with respect to the decrease in k. This benefits from the fact that OMC-ART uses key features to index images so that its computational cost during retrieval is low. Moreover, the hierarchical indexing structure of OMC-

ART essentially provides a batch-mode pre-ranking of the indexed images. Therefore, the groups of images similar to the query are likely to be selected for ranking prior to those of dissimilar images. This allows OMC-ART to stop the ranking process when the retrieval list is full and remains unchanged for a certain period of time.

To demonstrate the effective of the online indexing property of OMC-ART, we simulated the scenario of processing live streams of images and evaluated the processing times required by OMC-ART with online and offline learning to index the data set. Specifically, we separated the 15,000 images of NUS-WIDE data set into 15 groups of equal size and presented them sequentially to OMC-ART to build the indexing system. As shown in Figure 4, we observed that, with offline learning, the time cost of OMC-ART(offline) required to indexing the data set linearly increases with respect to the increase in the size of the data set. In contrast, that of OMC-ART(online) roughly remains the same as only the new data are handled.

6. CONCLUSIONS

This paper presented a novel idea for the automatic multimodal indexing and retrieval of weakly labeled image collections, wherein each image is associated with a textual description. In contrast to most existing approaches that aim to create a new feature space utilizing multimodal information for indexing images, the proposed OMC-ART aims to identify the representative features of each modality for groups of similar images and indexes the images using these key features. This idea is achieved by producing a two-layer hierarchical indexing structure for the images based on the heterogeneous data co-clustering algorithm, named GHF-ART. In addition, by extending GHF-ART with an adaptation method, OMC-ART is able to perform online learning, which favors web image collections requiring frequent update. With the proposed co-indexing method, OMC-ART allows flexible multimodal search by using either images, keywords, or a combination of both. Moreover, OMC-ART employs a carefully designed ranking algorithm, using which enables images more similar to the query to be more likely to be selected for ranking prior to those dissimilar ones. It also enables OMC-ART to stop the ranking process when the retrieval list is full and remains unchanged for a certain period of time.

In this paper, we have demonstrated the feasibility of the proposed co-indexing approach and the ranking algorithm. However, there remains places that require further exploration. First, OMC-ART requires the tuning of the vigilance parameters ρ for each modality to control the intra-cluster similarities. Currently they are manually tuned as discussed in Section 5.3, further efforts are required to make them self-adapted. Second, more effective methods for the key feature selection can be incorporated to enhance the indexing accuracy. Third, parallel implementation of OMC-ART can be studied in order to improve the efficiency of indexing very large data.

Acknowledgments

This research is supported by the National Research Foundation, Prime Minister's Office, Singapore under its IDM Futures Funding Initiative and administered by the Interactive and Digital Media Programme Office.

7. REFERENCES

[1] J. C. Caicedo, J. BenAbdallah, F. A. González, and O. Nasraoui. Multimodal representation, indexing, automated annotation and retrieval of image collections via non-negative matrix factorization. *Neurocomputing*, 76(1):50–60, 2012.

[2] J. C. Caicedo, J. G. Moreno, E. A. Niño, and F. A. González. Combining visual features and text data for medical image retrieval using latent semantic kernels. *In Proceedings of the international conference on Multimedia information retrieval*, pages 359–366, 2010.

[3] P. Chandrika and C. Jawahar. Multi modal semantic indexing for image retrieval. *In CIVR*, pages 342–349, 2010.

[4] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y. Zheng. NUS-WIDE: a real-world web image database from national university of singapore. *In Proceedings of the ACM international conference on image and video retrieval*, 2009.

[5] L. De Lathauwer, B. De Moor, and J. Vandewalle. A multilinear singular value decomposition. *SIAM journal on Matrix Analysis and Applications*, 21(4):1253–1278, 2000.

[6] P. Duygulu, K. Barnard, J. F. de Freitas, and D. A. Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. *In ECCV*, pages 97–112, 2002.

[7] H. J. Escalante, M. Montes, and E. Sucar. Multimodal indexing based on semantic cohesion for image retrieval. *Information Retrieval*, 15(1):1–32, 2012.

[8] Y. Gong, L. Wang, M. Hodosh, J. Hockenmaier, and S. Lazebnik. Improving image-sentence embeddings using large weakly annotated photo collections. *In Proceedings of the European Conference on Computer Vision (ECCV)*, pages 529–545, 2014.

[9] M. Li, X.-B. Xue, and Z.-H. Zhou. Exploiting multi-modal interactions: A unified framework. pages 1120–1125, 2009.

[10] R. Lienhart, S. Romberg, and E. Hörster. Multilayer pLSA for multimodal image retrieval. *In Proceedings of the ACM International Conference on Image and Video Retrieval*, 2009.

[11] T. Mei, Y. Rui, S. Li, and Q. Tian. Multimedia search reranking: A literature survey. *ACM Computing Surveys (CSUR)*, 46(3):38, 2014.

[12] L. Meng and A.-H. Tan. Semi-supervised hierarchical clustering for personalized web image organization. *In Proceedings of International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2012.

[13] L. Meng and A.-H. Tan. Community discovery in social networks via heterogeneous link association and fusion. *In Proceedings of the SIAM International Conference on Data Mining (SDM)*, pages 803–811, 2014.

[14] L. Meng, A.-H. Tan, and D. C. Wunsch. Vigilance adaptation in adaptive resonance theory. *In Proceedings of International Joint Conference on Neural Networks (IJCNN)*, pages 1–7, 2013.

[15] L. Meng, A.-H. Tan, and D. Xu. Semi-supervised heterogeneous fusion for multimedia data co-clustering. *IEEE Transactions on Knowledge and Data Engineering*, 26(9):2293–2306, 2014.

[16] Y. Mu, J. Shen, and S. Yan. Weakly-supervised hashing in kernel space. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3344–3351, 2010.

[17] L. Nie, M. Wang, Y. Gao, Z.-J. Zha, and T.-S. Chua. Beyond text QA: Multimedia answer generation by harvesting web information. *IEEE Transactions on Multimedia*, 15(2):426–441, 2013.

[18] L. Nie, M. Wang, Z.-J. Zha, G. Li, and T.-S. Chua. Multimedia answering: Enriching text QA with media information. *In SIGIR*, pages 695–704, 2011.

[19] A. W. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(12):1349–1380, 2000.

[20] J.-H. Su, B.-W. Wang, T.-Y. Hsu, C.-L. Chou, and V. S. Tseng. Multi-modal image retrieval by integrating web image annotation, concept matching and fuzzy ranking techniques. *International Journal of Fuzzy Systems*, 12(2):136–149, 2010.

[21] F. X. Yu, R. Ji, M.-H. Tsai, G. Ye, and S.-F. Chang. Weak attributes for large-scale image retrieval. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2949–2956, 2012.

[22] S. Zhang, M. Yang, X. Wang, Y. Lin, and Q. Tian. Semantic-aware co-indexing for image retrieval. *In Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1673–1680, 2013.