3-2013

Semi-supervised heterogeneous fusion for multimedia data co-clustering

Lei MENG

Ah-hwee TAN
Singapore Management University, ahtan@smu.edu.sg

Dong XU

Semi-Supervised Heterogeneous Fusion for Multimedia Data Co-clustering

Lei Meng, Ah-Hwee Tan, *Senior Member, IEEE* and Dong Xu, *Member, IEEE*

Abstract—Co-clustering is a commonly used technique for tapping the rich meta-information of multimedia web documents, including category, annotation, and description, for associative discovery. However, most co-clustering methods proposed for heterogeneous data do not consider the representation problem of short and noisy text and their performance is limited by the empirical weighting of the multi-modal features. In this paper, we propose a generalized form of Heterogeneous Fusion Adaptive Resonance Theory, called GHF-ART, for co-clustering of large-scale web multimedia documents. By extending the two-channel Heterogeneous Fusion ART (HF-ART) to multiple channels, GHF-ART is designed to handle multimedia data with an arbitrarily rich level of meta-information. For handling short and noisy text, GHF-ART does not learn directly from the textual features. Instead, it identifies key tags by learning the probabilistic distribution of tag occurrences. More importantly, GHF-ART incorporates an adaptive method for effective fusion of multi-modal features, which weights the features of multiple data sources by incrementally measuring the importance of feature modalities through the intra-cluster scatters. Extensive experiments on two web image data sets and one text document set have shown that GHF-ART achieves significantly better clustering performance and is much faster than many existing state-of-the-art algorithms.

Index Terms—Semi-supervised learning, heterogeneous data co-clustering, multimedia data mining.

1 Introduction

The increasingly popularity of social networking websites, such as Flickr and Facebook, has led to the explosive growth of multimedia web documents sharing online. In order to provide easy access for users to browse and manage large-scale repositories, effective organization of those documents with common subjects is desired. Clustering techniques, designed to identify groupings of data in multi-dimensional feature space based on measured similarity, are often applied to this task. As web multimedia resources are often attached with rich meta-information, for example, category, annotation, description, images and surrounding text, how to utilize the additional information to enhance the clustering performance poses a challenge to traditional clustering techniques.

In the recent years, the heterogeneous data co-clustering approach, which advances from the clustering of one data type to the co-clustering of multiple data types, has drawn much attention and been applied to the image and text domains [1], [2], [3], [4], [5]. However, the algorithms follow the similar idea of linearly combining the objective functions of each feature modality and subsequently minimizing the global cost. For the co-clustering of multimedia data, existing algorithms face three challenges elaborated as follows. Firstly, similar to the short text clustering problem [6], meta-information is usually very short and therefore the extracted tags cannot be effectively weighted by traditional data mining techniques such as term frequency-inverse document frequency (tf-idf). Secondly, the weights of features in the objective function still rely on empirical settings, which usually leads to a sub-optimal result. Finally, this approach requires an iterative process to ensure the convergency, which leads to high computational complexity. Thus, existing methods are only applicable to small data sets consisting of up to a thousand of documents but become very slow and not scalable to big data.

In view of the above issues, a self-organizing neural network called Heterogeneous Fusion Adaptive Resonance Theory (HF-ART) [7] has been recently proposed for web image co-clustering, which performs fusion of visual and textual features as a mapping across two feature spaces. HF-ART achieves effective representation of the surrounding text by modeling the cluster prototype of textual features using probabilistic distribution of tag occurrences, and addresses the problem of feature weighting by employing a robustness measure to weight the features by learning from the intra-cluster scatters. Moreover, HF-ART is semi-supervised as it is able to take in prior knowledge by initializing the network with pre-defined clusters, indicating regions of interests to users. Different from traditional semi-supervised clustering techniques such as [4], in which the user-provided knowledge is rarely reflected by the resulting clusters, HF-ART can incrementally generalize and preserve the learnt knowledge by identifying and learning from relevant input patterns, and present the resulting clusters, reflecting user preferences, directly to the users.

Whereas HF-ART is restricted to two pattern channels, in this paper, we propose a generalized heterogeneous data co-clustering algorithm, termed Generalized Heterogeneous Fusion Adaptive Resonance Theory (GHF-ART), for fast and robust web multimedia data co-

Manuscript received Aug. 2, 2012; revised Dec. 23, 2012.
Lei Meng, Ah-Hwee Tan and Dong Xu are with the School of Computer Engineering, Nanyang Technological University, 50 Nanyang Avenue, Singapore 639798. E-mail: {meng0027, asahtan, dongxu}@ntu.edu.sg.

clustering. By extending HF-ART from a two-channel model to multiple feature channels wherein each channel may receive different types of data patterns, GHF-ART is designed to handle multimedia data with an arbitrarily rich level of meta-information. Accordingly, the adaptive feature weighting algorithm has also been generalized by evaluating a robustness measure for each of the multiple feature channels.

The performance of GHF-ART has been evaluated on two public web image data sets, namely the NUS-WIDE [8] and Corel data sets, and a public text document set, known as the 20 Newsgroups data set [9]. Our empirical results show that GHF-ART consistently achieves better cluster quality and is much faster than many state-of-the-art heterogeneous data co-clustering algorithms.

The rest of this paper is organized as follows. Section 2 reviews related works on image-text fusion and heterogeneous data co-clustering. The problem formulation of heterogeneous data co-clustering are described in section 3. An introduction of Heterogeneous Fusion ART is presented in section 4. The details of GHF-ART are introduced in section 5. The experiment results are presented in section 6. The last section summarizes our work and highlights the future work.

2 Related Work

2.1 Image-Text Fusion

Information fusion aims to process multiple interrelated data modalities in a unified way and identify their underlying interactions. Along with the rich multimedia contents published on the World Wide Web, information fusion has became an essential technique on various applications, such as multi-document summarization [10], [11] and multi-modal multimedia indexing and retrieval [12], [13], [14], [15].

Our work is related to the fusion of visual and textual features through clustering methods. In the early works, visual and textual features are concatenated into a single vector [16] or used in a consequence manner [17]. However, the first approach usually cannot achieve desired results since the concatenated features come from different resources and cannot well represent the key features of documents. Besides, the second method, which use textual and visual features consequently to generate a two-layer cluster structure, suffers from the problem of error propagation and the usage of visual features in the second step has no contribution to improve the clustering quality. Jiang et al. [18] interpret the fusion of visual and textual features as identifying pairs of related images and texts, and propose two methods for learning the image-text associations. The first method is based on the vague transformation [19] that models the associations between images and texts by measuring the visual-textual similarities. The other method is based on Fusion ART [20], which incrementally learns a set of prototypical image-text pairs from the data set.

A large literature of recent works are based on graph theory. Gao et al. [21] propose a Consistent Bipartite Graph Co-partitioning (CBGC), which interprets the image-text co-clustering task as a tripartite graph and transforms the partitioning of the tripartite graph into the simultaneous partitioning of the visual and textual sub-graphs. In this way, CBGC models the solution as a multi-objective optimization problem which is solved by the Semi-Definite Programming (SDP). A similar work Consistent Isoperimetric High-order Co-clustering (CIHC) [2] also considers the co-clustering problem as the partitioning of a tripartite graph. Different from CBGC, CIHC models the problem by an extended Isoperimetric Co-clustering Algorithm (ICA) [22] which is solved by a sparse system of linear equations. Cai et al. [23] propose a Multi-modal Spectral Clustering (MMSC) which uses a unified objective function to iteratively optimize the clustering results of each feature modality and their combination. In [24], a Multi-modal Constraint Propagation (MMCP) is proposed which first defines the random walk on multiple graphs and then deduces the results by quadratic optimization method. However, it requires to set many empirical parameters settings, such as the prior graph probabilities and the number of clusters, which are usually inapplicable to large-scale data set.

2.2 Heterogeneous Data Co-clustering

Heterogeneous data co-clustering approach addresses the problem of simultaneously integrating multiple types of data for clustering. Typically, the primary data, i.e. the documents, and the associative information, i.e. the attached meta-information, are modeled into star-structured relational data [4] and the co-clustering task is to find an optimal clustering of the documents according to all types of features. Considering different model formulation, existing algorithms can be categorized into three categories: models based on graph theory, Non-negative Matrix Factorization and information theory.

Graph theory based approach is widely used for the co-clustering task. Gao et al. [1] generalize their prior work on image-text co-clustering [21] for heterogeneous data co-clustering. Long et al. [3] propose a graph-based model, Spectral Relational Clustering (SRC), which first introduce a collective clustering based on minimizing the reconstruction error of both object affinity matrix and feature matrix, and then derive an iterative spectral clustering algorithm accordingly for the factorization of these relational matrices. However, SRC requires solving the eigen-decomposition problem which is inefficient for large-scale data sets. In addition, a separate clustering algorithm (in this case K-means) is used to obtain the final clustering. This is the common drawback of many graph theoretical clustering algorithms.

The Non-Negative Matrix Factorization (NMF) approach has been applied for text document clustering [25] and extended to a Co-clustering framework [26]. In the recent years, NMF has been extended to the

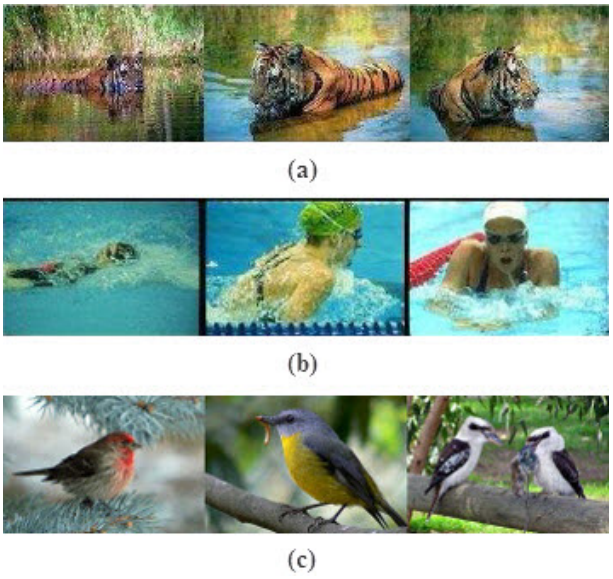

Fig. 1: Examples of web images consistent in both image content and high-level semantics.

multi-modal co-clustering task. Chen et al. [4] propose a symmetric non-negative matrix tri-factorization algorithm, called Semi-Supervised NMF (SS-NMF), which minimizes the global reconstruction error of all the relational matrices of the central type data and features. This method can derive a latent semantic space which reveals the relations between each data item and a pre-defined number of clusters (the axis). The cluster membership of each data item is determined by the largest projection value among all clusters. Moreover, by incorporating user-provided constraints, SS-NMF can derive new relational matrices through distance learning algorithm to enhance the clustering performance.

Bekkerman et. al [27] propose the Combinatorial Markov Random Fields (Comrafs) for the multi-modal information co-clustering based on the information Bottleneck theory and apply it in various fields such as semi-supervised learning [28], image clustering [5] and cluster analysis [29]. Comrafs constructs the Markov Random Fields wherein each modality of data is modeled as a combinatorial random variable which take values from all the possible partitions, and the edges between all variables are represented by Mutual Information. The clustering process is inferred by an information-theoretic objective. One potential problem of this approach is the time complexity. As Comrafs needs to traverse all subsets of the data samples for each data modality, the computational complexity will increase significantly with the increase in the size of data set.

3 Problem Formulation

Considering a set of documents $\mathcal{D} = \{doc_n|_{n=1}^N\}$ with the associated meta-information, which may be tags, category information and surrounding text, each document doc_n may be represented by a multi-channel input pattern $\mathcal{I} = \{\mathbf{x}^k|_{k=1}^K\}$, where $\mathbf{x}^k$ is a feature vector extracted from the document or one type of meta-information. The goal of the heterogeneous data co-clustering task, as defined in this paper, is to partition the set of N documents into a set of clusters $\mathcal{C} = \{c_j|_{j=1}^J\}$ by evaluating the similarity between the input patterns of the documents according to their corresponding feature vectors such that the documents belonging to the same cluster should be more similar to each other than to the documents of the other clusters. For example, in the image domain, the co-clustering task may be to identify similar images according to both the visual content and the surrounding text. In each cluster, the images therein are similar in image content and the high-level semantics reflected from the image content are consistent. Similarly, in the text domain, the co-clustering task is to consider both the features of the text document and the meta-information, such as category information and authors.

As reviewed in the previous section, the heterogeneous data co-clustering task presents a number of issues and challenges, especially for multimedia data set. We discuss the key challenges in three aspects as follows.

1) **Representation of document content:** The representation issue of text documents has been well studied in literature. Typically, text documents are represented by the keywords appearing in the document collection, each of which is weighted based on its frequency in and cross the documents, known as tf-idf. On the other hand, visual representation of images is still a challenge nowadays. Current techniques for visual feature extraction are based on color histogram, edge detection, texture orientation and scale-invariant points so that the visual features are inadequate to represent the images at the semantic level, a problem known as semantic gap. It leads to difficulties to group the images with very different appearance (Fig. 1(c)) or to distinguish those with similar background (Fig. 1(a) and 1(b)).
2) **Representation of meta-information:** The meta-information of documents provides additional knowledge which indicates the relations between documents from another perspective. However, in both image and text domains, the problem of noisy tags exists. Specifically, although the extracted tags from the meta-information of documents usually contain the key tags that are helpful for identifying the correct groupings of documents, a large number of noisy tags exist which contribute nothing or even indicate incorrect relations between documents. How to identify key tags from noisy text is also an open problem in tag ranking [30], [31].
3) **Integrating multiple types of features:** It is the key challenge which is related to heterogeneous data utilization for clustering. Existing works, described in Section 2, typically rely on some global optimization methods for the partitioning of each feature modality. However, they do not address

the problem of weighting the feature modalities in their objective functions. Instead, either a uniform weighting or some empirical settings are used, which may not yield the desirable results.

4 Heterogeneous Fusion ART

Adaptive Resonance Theory (ART) [32] is a neural theory of cognitive information processing. ART performs unsupervised learning by modeling clusters as memory prototypes and encodes each input pattern incrementally through a two-way similarity measure, which simulates how human brain capture, recognize and memorize information of objects and events. As long as the difference between the input pattern and the selected prototype does not exceed a threshold called vigilance parameter, the input pattern is considered a member of the selected cluster. ART takes the advantages of fast and stable learning as well as the incremental manner, and has shown strong noise immunity [33].

By extending ART from a single input field to multiple ones, Fusion ART [20] provides a general architecture for simultaneously learning of multi-modal feature mappings. Specifically, Fusion ART performs real-time search for suitable clusters and learns to encode the mappings of multi-modal features in an incremental manner. A previous work [34] shows the viability of Fusion ART for integrating visual and textual features for image-text co-clustering. However, its performance is limited by the textual feature representation and learning.

The architecture of HF-ART is a two-channel Fusion ART. Different from Fusion ART, HF-ART employs heterogeneous learning for the features of documents and meta-information respectively to achieve effective cluster prototypes. The proposed learning method models the corresponding cluster prototype by the probability distribution of tag occurrences, which helps to address the problem of noisy tags, especially for data type of which insufficient statistic information is provided. Besides, by employing the robustness measure, the contribution parameter is adaptively adjusted by learning the inner-class scatter of clusters. In this way, the problem of weighting multi-modal features is solved by learning from the cluster structure of input patterns.

5 Generalized Heterogeneous Fusion ART

Generalized Heterogeneous Fusion ART (GHF-ART) (Fig. 2) extends the HF-ART from two channels to multiple channels so that GHF-ART can be applied to the clustering of more than two modalities wherein each channel may receive different types of data patterns. More importantly, by generalizing the feature construction methods for multimedia documents and incorporating an adaptive channel weighting algorithm, GHF-ART is able to effectively integrate different types of features across multiple pattern channels. Whereas most current

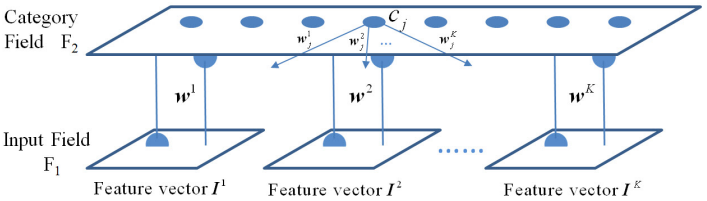

Fig. 2: The Architecture of Generalized Heterogeneous Fusion ART.

works [2], [3], [4], [5] employ statistical methods, the proposed GHF-ART model performs the heterogeneous data co-clustering using a self-organizing neural network. In essence, GHF-ART simultaneously learns the multi-dimensional mappings across multiple feature spaces to the category space. The clustering process of GHF-ART thus partitions the category space into regions of clusters by incrementally learning the cluster prototypes from the input patterns and identifying the key features.

Moreover, with the incremental characteristics of Adaptive Resonance Theory, GHF-ART may perform semi-supervised learning by taking in the user preferences in the form of prior knowledge to initialize the cluster structure before clustering. Essentially, the user may identify groupings of documents wherein documents in the same group are deemed to be similar to each other. During the network initialization step, GHF-ART first generates clusters for these user-specified groups of documents. These pre-defined clusters can then be treated as user-defined projection from the feature space to the category space. During the subsequent clustering process, these user-defined clusters can be further generalized by recognizing and learning from similar input patterns, while new clusters can still be created automatically for novel patterns dissimilar to existing clusters. By incorporating user preferences, the predefined clusters help to construct better cluster structure comparing to one using pure data driven clustering.

The dynamics of GHF-ART algorithm is summarized as follows.

Input vectors: Let $\mathcal{I} = \{\mathbf{x}^k|_{k=1}^K\}$ denote the multi-channel input pattern, where $\mathbf{x}^k$ is the feature vector for the k-th feature channel. Note that, with complement coding [35], $\mathbf{x}^k$ is further augmented with a complement vector $\bar{\mathbf{x}}^k$ such that $\bar{\mathbf{x}}_i^k = 1 - \mathbf{x}_i^k$ in the input field F_1.

Weight vectors: Let $\{\mathbf{w}_j^k|_{k=1}^K\}$ denote the weight vectors associated with the j-th cluster c_j in the category field F_2.

Parameters: The GHF-ART's dynamics is determined by choice parameter $\alpha > 0$, learning parameter $\beta \in [0, 1]$, contribution parameters $\gamma^k \in [0, 1]$ and vigilance parameters $\rho^k \in [0, 1]$ for $k = 1, ..., K$.

The clustering process of GHF-ART comprises four key steps: 1) network initialization: if the user preferences are provided, generate a cluster for each group of documents. Specifically, each cluster c_j has K weight vectors $\{\mathbf{w}_j^k|_{k=1}^K\}$, obtained by averaging the values of

the feature vectors of the documents it contains. Otherwise, generate an uncommitted cluster with all the feature values of its weight vectors set to 1; 2) category choice: for each input pattern $\mathcal{I} = \{\mathbf{x}^1, \ldots, \mathbf{x}^K\}$, select the most suitable cluster (winner cluster) c_{j^*}, which has the maximum score calculated by a choice function $T(c_j, \mathcal{I})$ $(j = 1, \ldots, J)$; 3) template matching: evaluate the similarity between the input pattern $\mathcal{I}$ and the winner c_{j^*} using a match function $M(c_{j^*}, \mathcal{I})$ and a vigilance parameter ρ. If the winner satisfies the vigilance criteria, a resonance occurs which leads to the learning step. Otherwise, a new winner is selected from the rest of the clusters in the category field. If no winner satisfies the vigilance criteria, a new cluster is generated to encode the input pattern; and 4) prototype learning: if c_{j^*} satisfies the vigilance criteria, its corresponding weight vectors $\mathbf{w}_{j^*}^k$ $(k = 1, \ldots, K)$ are updated through a learning function (see Section 5.3). The algorithm stops when all the input patterns are presented.

5.1 Feature Extraction

5.1.1 Feature Extraction for Document Content

In our work, a document can be either an image or an article. For an image, the feature vector is the concatenation of multiple types of visual features. For an article, we extract the term frequency-inverse document frequency (tf-idf) features. Since the ART-based algorithm requires the values of input in the interval [0,1], we further apply min-max normalization on the features.

5.1.2 Feature Extraction for Meta-Information

As the meta-information (e.g. the surrounding text for a web image or the author information for an article) is usually short and noisy, traditional text mining techniques cannot effectively weight the tags. For example, the tf-idf features usually leads to feature vectors with a flat distribution of low values [6]. Therefore, we model the textual features to indicate the presence of tags such that the probabilistic distribution of tags occurrences in the given clusters can be subsequently learnt as the cluster prototype of textual features through the proposed learning function (9).

We construct the textual feature vector for the meta-information based on a textual table consisting of all distinct tags in the whole image set expressed by $\mathcal{G} = \{g_1, \ldots, g_M\}$. Then, we denote the textual feature vector for the n-th document doc_n as $\mathbf{t}_n = [t_n^1, \ldots, t_n^M]^\top$, where t_n^m corresponds to the m-th tag g_m in $\mathcal{G}$. The value of t_n^m is given by:

$$t_n^m = \begin{cases} 1, & if\ g_m \in doc_n \\ 0, & otherwise \end{cases}. \tag{1}$$

The feature vector indicates a point in the textual feature space of M dimensions constructed by all tags. Therefore, more common tags in two given images lead to a shorter distance in the feature space of the GHF-ART.

5.2 Similarity Measure

We adopt the similarity measure of Fusion ART [20] to select the best matching cluster for the input pattern. Considering a document doc_n with its corresponding multi-channel input pattern $\mathcal{I} = \{\mathbf{x}^1, ..., \mathbf{x}^K\}$, the cluster selection process consists of two stages, namely category choice and template matching. In the first step, a choice function is applied to evaluate the overall similarity between the input pattern and the template pattern of each cluster in the category field. Specifically, the choice function for each cluster c_j is defined by

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

where the fuzzy AND operation $\wedge$ is defined by $(\mathbf{p} \wedge \mathbf{q})_i \equiv \min(\mathbf{p}_i, \mathbf{q}_i)$, and the norm $|.|$ is defined by the ℓ_1 norm.

After identifying the cluster having the highest value as the winner c_{j^*}, we use a match function to evaluate if the similarity between the input pattern $\mathcal{I}$ and the winner c_{j^*} meets the vigilance criteria. The match function, for the k-th feature channel, is defined by

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

If, for all the K feature channels, the corresponding match function satisfies the vigilance criteria $M(c_{j^*}, \mathbf{x}^k) > \rho^k$ $(k = 1, ..., K)$, a resonance occurs and the input pattern is categorized into the winner cluster. Otherwise, a reset occurs to select a new winner from the rest of the clusters in the category field.

PROPERTY 1. *Using the category choice and template matching functions, each input pattern is categorized into the cluster with the best matching feature distribution.*

Proof: From (2), we observe that, for each feature channel k, the similarity is calculated by the ratio of the intersection $|\mathbf{x}^k \wedge \mathbf{w}_j^k|$ and the corresponding cluster prototype $|\mathbf{w}_j^k|$. If we interpret the feature vector using histogram, the most similar feature distribution produces the largest value of $\frac{|\mathbf{x}^k \wedge \mathbf{w}_j^k|}{\alpha + |\mathbf{w}_j^k|}$. Taking into account all of the feature channels, the choice function measures the overall similarity between the input pattern $\mathcal{I}$ and the cluster c_j across all of the K feature channels. Thus, the category choice procedure selects the cluster whose feature distribution across all features is the most satisfied by the input pattern.

Subsequently, the template matching procedure defined by (3) evaluates if the selected winner matches well with the feature distribution of the input pattern, controlled by the vigilance parameter ρ^k. With a reasonable setting of ρ^k, the clusters that do not match the feature distribution of the input pattern are rejected.

If all the existing categories are not fit for the input pattern, a new cluster is generated and the prototypes are set by the features of the input pattern. In this

way, each input pattern will be grouped into the best matching cluster. □

5.3 Learning strategies for multi-modal features

5.3.1 Learning key features of document content

We use the learning function of Fusion ART [20] to learn the cluster prototype for the document content. Given an input document with its multi-channel input pattern $\mathcal{I} = \{\mathbf{x}^1, \ldots, \mathbf{x}^K\}$ and the winner cluster c_{j^*}, if x^k is the feature vector for document content, then the learning function for the corresponding weight vector $\mathbf{w}_{j^*}^k$ is defined by

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

PROPERTY 2. *The learning function defined by (4) incrementally identifies the key features from the input patterns.*

Proof: The learning function defined by (4) consists of two components $\mathbf{x}^k \wedge \mathbf{w}_{j^*}^k$ and $\mathbf{w}_{j^*}^k$, in which the first component is the intersection between the input pattern and the cluster prototype and the second one is the cluster prototype. We observe that whatever the value of the learning rate β is, the values of the new cluster prototype, for each component of the feature vector, will not exceed the old one. That is, if the components of the feature vector is unstable in values, the prototype learns a small value. In this way, the cluster prototype learns from the input pattern by stably depressing the rarely high and unstable components while preserving the key and frequently high ones. □

5.3.2 Learning key features of meta-information

Directly learning the key tags of clusters from individual documents represented by traditional weighting techniques is usually biased by the limited tag lexicon and statistical information. Based on the above consideration, we propose to model the cluster prototype of textual features by the probabilistic distribution of tag occurrences. In this way, the weights of noisy tags are depressed while the key and sub-key tags can be preserved.

Assuming the winner c_{j^*} contains L documents, denoted as $c_{j^*} = \{doc_1, \ldots, doc_L\}$. Recall that in Section 5.1.2, we denote the feature vector for the meta-information of doc_l as $\mathbf{t}_l = [t_l^1, \ldots, t_l^M]^\top$, so the weight vector for the k-th feature channel of cluster c_{j^*} can be represented as $\mathbf{w}_{j^*}^k = [w_{j^*,1}^k, \ldots, w_{j^*,M}^k]^\top$. Then, the probability of occurrences of the m-th tag in $\mathcal{G}$ in the winner cluster c_{j^*} having L documents is calculated by:

$$w_{j,m}^k = p_L(g_m|c_{j^*}) = \frac{\sum_{l=1}^{L} t_l^m}{L}. \tag{5}$$

Therefore the prototype for the textual features of cluster c_{j^*} can be represented by

$$\mathbf{w}_j^k = [p_L(g_1|c_{j^*}), \ldots, p_L(g_M|c_{j^*})]^\top. \tag{6}$$

Now we introduce the sequential factor. We treat $p_L(g_m|c_{j^*})$ in (5) as the state for time L. Assuming a new document doc_{L+1} is grouped into cluster c_{j^*}, we derive the relationship between the probabilities of occurrence of the m-th tag at time L and $L+1$ by

$$p_{L+1}(g_m|c_{j^*}) = \frac{\sum_{l=1}^{L+1} t_l^m}{L+1} = \frac{L}{L+1} p_L(g_m|c_{j^*}) + \frac{t_{L+1}^m}{L+1}. \tag{7}$$

Therefore, the general form of learning function for $w_{j^*,m}^k$ is defined by

$$\hat{w}_{j^*,m}^k = \frac{L}{L+1} w_{j^*,m}^k + \frac{t_{L+1}^m}{L+1}. \tag{8}$$

Considering t_{L+1}^m equals to either 0 or 1, we further simplify the learning function for $\mathbf{w}_{j^*}^k = [w_{j^*,1}^k, \ldots, w_{j^*,M}^k]^\top$ such that

$$\hat{w}_{j^*,m}^k = \begin{cases} \eta w_{j^*,m}^k, & if\ t_{L+1}^m = 0 \\ \eta(w_{j^*,m}^k + \frac{1}{L}), & otherwise \end{cases}. \tag{9}$$

where $\eta = \frac{L}{L+1}$.

5.4 Self-adaptive Parameter Tuning

The settings of vigilance parameter ρ and contribution parameter γ affect the clustering results greatly. Using some fixed values will certainly limit the robustness of GHF-ART for a diverse range of data sets. Therefore, self-adaptive tuning of the two parameters is desirable.

5.4.1 Match Tracking Rule

The original match tracking rule was first used in ARTMAP [36] to maximize generalization with a minimum number of cluster nodes. GHF-ART utilizes a generalized form of match tracking rule, wherein the vigilance value of each feature channel can be adapted.

At the beginning of each input pattern presentation, the vigilance parameters of all feature channels $\{\rho^1, ..., \rho^K\}$ are set to a baseline ρ_0. A change in the vigilance values is triggered when the template matching process causes a reset. The process is formalized as:

$$\hat{\rho}^k = M(c_{j^*}, \mathbf{x}^k) + \varepsilon. \quad (k = 1, ..., K) \tag{10}$$

where $\varepsilon > 0$ is a very small value and $M(c_{j^*}, \mathbf{x}^k)$ is defined as in (3).

5.4.2 Robustness Measure of Features

The contribution parameter specifies the weighting factor given to each feature channel during the category choice process. Intuitively, the feature channel which is more robust in distinguishing the classes of the patterns should have a higher weight. Therefore, we want to scale the robustness of the feature channels by learning from the input patterns rather than following an empirical setting. In view that a robust feature channel represents the documents belonging to the same class stably, namely with a small scatter in the cluster weights, it can be measured by the difference between the intra-cluster patterns and the cluster prototypes (weights). Consider a cluster c_j and the intra-cluster documents

$\{doc_1,\ldots, doc_L\}$. By denoting the features vectors of doc_l as $\mathcal{I}_l = \{\mathbf{x}_l^1,..., \mathbf{x}_l^K\}$ and the weight vectors of the cluster c_j as $\mathbf{w}_j = \{\mathbf{w}_j^1, ..., \mathbf{w}_j^K\}$, we define the Difference for the k-th feature vector in c_j as follows:

$$D_j^k = \frac{\frac{1}{L}\sum_l |\mathbf{w}_j^k - \mathbf{x}_l^k|}{|\mathbf{w}_j^k|}. \tag{11}$$

Subsequently, the overall difference of one feature vector can be evaluated by averaging the difference of all clusters, defined by:

$$D^k = \frac{1}{J}\sum_j D_j^k, \tag{12}$$

where J is the number of clusters. Therefore, the robustness of the k-th feature modality can be measured by

$$R^k = \exp(-D^k). \tag{13}$$

When D^k is 0, R^k becomes 1, which means that this feature can well represent the images belonging to one class. In contrast, when D^k is very large, R^k approaches zero. The expression implies that the feature with higher difference is not robust and has lower reliability. Thus, in a normalized form, the contribution parameter γ for the k-th feature channel can be expressed by

$$\gamma^k = \frac{R^k}{\sum_{k=1}^K R^k}. \tag{14}$$

This equation shows the rule for tuning the contribution parameter during the clustering process. Initially, the contribution parameter is given by equal weights based on the intuition that the powers of all features are the same. Subsequently, the value of γ changes along with the encoding of input patterns.

The tuning of contribution parameters occurs after each resonance, i.e. the clustering epoch for each input pattern, which can be computationally expensive. For efficiency purpose, we further derive a method to incrementally update the contribution parameter values, according to the learning functions defined in (4) and (9). We consider the update equations in two cases:

- **Resonance in existing cluster**: Assuming the input pattern is assigned to an existing cluster c_j. In this case, only the change of D_j^k should be considered. For the k-th feature channel, the update equations for document content and meta-information are defined by (15) and (16) respectively:

$$\hat{D}_j^k = \frac{\eta}{|\hat{\mathbf{w}}_j^k|}(|\mathbf{w}_j^k|D_j^k + |\mathbf{w}_j^k - \hat{\mathbf{w}}_j^k| + \frac{1}{L}|\hat{\mathbf{w}}_j^k - \mathbf{x}_{L+1}^k|) \tag{15}$$

$$\hat{D}_j^k = \frac{\eta}{|\hat{\mathbf{w}}_j^k|}(\eta D_j^k + |\hat{\mathbf{w}}_j^k - \eta\mathbf{w}_j^k| + \frac{1}{L}|\hat{\mathbf{w}}_j^k - \mathbf{x}_{L+1}^k|). \tag{16}$$

  After the update for all of the feature channels, the new contribution parameter can then be obtained by calculating (12)-(14). In this way, the computational complexity reduces from $O(n_i n_f)$ to $O(n_f)$, where n_f denotes the dimension of the feature channels and n_i denotes the number of documents.
- **Generation of new cluster**: When generating a new cluster, the differences of other clusters remain unchanged. Therefore, it just introduces a proportionally change of the robustness. Considering the robustness R^k $(k = 1, ..., K)$ for all of the feature channels, the update equation for the k-th feature channel is derived as:

$$\hat{\gamma}^k = \frac{\hat{R}^k}{\sum_{k=1}^K \hat{R}^k} = \frac{(R^k)^\eta}{\sum_{k=1}^K (R^k)^\eta}, \tag{17}$$

5.5 Summary of GHF-ART algorithm

The complete algorithm of GHF-ART is summarized as follows.

Clustering algorithm of GHF-ART

1) Generate pre-defined clusters for the initial network based on user preferences. If no prior knowledge is received, create an uncommitted cluster with all weight vectors containing 1's.
2) For each document, present its corresponding input pattern $\mathcal{I} = \{\mathbf{x}^1, \ldots, \mathbf{x}^K\}$ into the input field F_1.
3) For each cluster c_j in the category field F_2, calculate the choice function $T(c_j, \mathcal{I})$ defined in (2).
4) Identify the winner c_{j^*} with the largest value of the choice function such that $j^* = \arg\max_{j:c_j \in F_2} T(c_j, \mathcal{I})$.
5) Calculate the match function $M(c_{j^*}, \mathbf{x}^k)$ $(k = 1, ...K)$ defined in (3).
6) If $\exists k$ such that $M(c_{j^*}, \mathbf{x}^k) < \rho^k$, set $T(c_{j^*}, \mathcal{I}) = 0$, update ρ^k $(k = 1, ..., K)$ according to (10), go to 4; else, go to 7.
7) If the selected c_{j^*} is uncommitted, set each cluster prototype to the corresponding feature vector of the input pattern such that $\mathbf{w}_{j^*}^k = \mathbf{x}^k$ $(k = 1, ..., K)$, and update γ according to (17) and create a new uncommitted node, go to 9; else, go to 8.
8) Update $\mathbf{w}_{j^*}^k$ $(k = 1, ..., K)$ according to (4) and (9) respectively and update γ according to (12)-(16).
9) If no input pattern exist, algorithm stops. Otherwise, go to 2.

5.6 Time Complexity

The time complexity of GHF-ART depends on the search of suitable categories and the update of contribution parameter. The first step calculates the choice and match function defined in (2) and (3), which is $O(n_c n_f)$, where n_c denotes the number of clusters and n_f denotes the number of feature dimension of both visual and textual features. The second step contains two cases: 1) the input pattern is grouped into one of existing clusters; and 2) a new cluster is generated for the input pattern. For the first case, the new contribution parameter is calculated by (12)-(16). The time complexity of (15) and (16) is $O(n_f)$ and that of (12)-(14) is $O(1)$. For the second case, the contribution parameter is updated according to (17), whose time complexity is $O(1)$. Assuming there are n_i input patterns, the overall time complexity is $O(n_i n_c n_f)$.

In comparison, the time complexity of CIHC co-cluster algorithm is $O(QR\{n_i n_f\} + (n_i + n_f) log^{(n_i+n_f)})$, where QR{.} is the time for QR matrix decomposition. The time complexity of NMF is $O(tn_c n_i n_f)$, SRC is $O(t(max(n_i^3, n_f^3) + n_c n_i n_f))$, Comrafs is $O(t(max(n_i^3, n_f^3))$, where t is the number of iterations in the algorithm. We observe that GHF-ART requires the least time cost and maintains a linear increase of running time with the increase of the data set.

6 Experiments

6.1 NUS-WIDE Data Set

The NUS-WIDE data set [8] is the largest well-annotated web image set with filtered surrounding text, which consists of 269,648 images and 81 concepts as ground-truth. The images are downloaded from the famous photo sharing website *Flickr.com*. To evaluate the clustering performance of our method on large scale image sets, we collect a total of 23,284 images belonging to nine biggest classes of NUS-WIDE data set, including dog, bear, cat, bird, flower, lake, sky, sunset and wedding, each of which contains nearly 3000 images, except bear (1,271 images) and wedding (1,353 images).

We utilize the visual content and surrounding text of images for clustering. For the visual features, we use a concatenation of Grid Color Moment (255 features), Edge Direction Histogram (73 features) and Wavelet Texture (128 features). We use the above three types of global features as they can be efficiently extracted and have been shown to be effective for image content representation [8]. Finally, each image is represented as a vector of 426 features. We construct the texture feature vector by considering all distinctive and high frequency tags in the surrounding text of images. After filtering the infrequency tags, we have a total of 1,142 textual features and each image is associated with seven tags on average.

6.1.1 Performance of Robustness Measure

In the experiments, we set the choice parameter $\alpha = 0.01$, the learning parameter $\beta = 0.6$ and the baseline vigilance parameter $\rho_0 = 0.1$. Small choice parameter of $\alpha = 0.01$ is commonly used as it has been shown that the clustering performance is generally robust to this parameter [37]. We empirically use $\beta = 0.6$ to tune the cluster weight towards the geometric center of the cluster. In our experiments, the performance of GHF-ART remains roughly the same when the learning parameter changes from 0.3 to 0.8. In view that the vigilance parameter has a direct effect on the number of generated clusters, we use $\rho_0 = 0.1$ which produces a small number of small clusters containing less than 1% of the data patterns. In our experiments, we find that the performance of GHF-ART improves significantly when ρ increases to 0.1. Beyond that, the performance improvement is rather small but the number of clusters increase almost linearly. Therefore, we use $\rho_0 = 0.1$ consistently in all our experiments. Other vigilance values

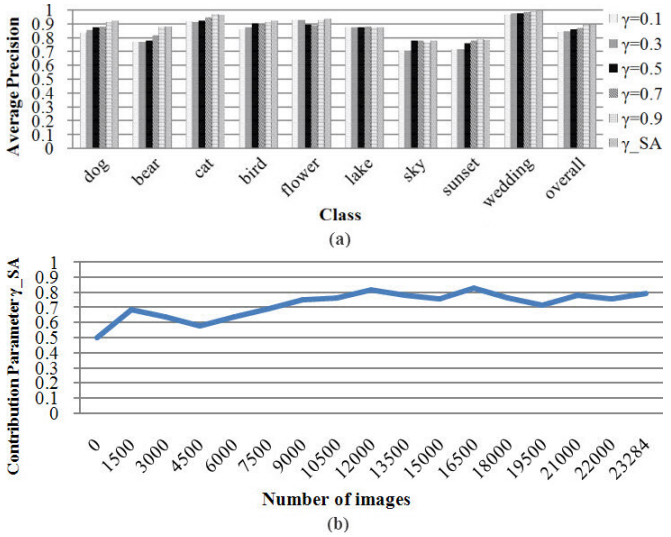

Fig. 3: (a) Clustering performance using fixed contribution parameters (γ) and self-adapted contribution parameter (γ_{SA}); (b) Tracking of γ_{SA} of textual feature channel on NUS-WIDE data set.

may still work, but a higher vigilance value may lead to a better performance in precision but may create many more clusters resulting in poorer generalization.

We evaluate the performance of robustness measure by comparing the clustering performance of GHF-ART using the self-adapted contribution parameter γ_{SA} with that of fixed values. Since we utilize two channels for visual and textual features respectively, we vary the contribution parameter of textual features and calculate that of visual features by (14). The result is shown in Fig. 3(a). We observe that, without prior knowledge, the self-adaptive tuning method always has comparable performance with the best settings and even slightly improve the results in several classes. The average precision across all classes shows that the overall performance of the robustness measure is slightly better than the best results of the fixed settings of the contribution parameter. Besides, the time cost of GHF-ART with fixed settings is 9.610 seconds and that with the robustness method is 9.832 seconds. Therefore, this method is effective and efficient for solving the tuning problem of contribution parameter and is also scalable to big data.

To understand how the robustness measure works, we show the value tracking of γ_{SA} of the textual feature channel in Fig. 3(b). We observe that, despite the initial fluctuation, the value of γ_{SA} climbs from 0.5 to 0.8 and then stabilizes in the interval of [0.7, 0.8]. The initial fluctuation should due to the order of input pattern presentation. As the robustness measure adjusts the contribution parameters along with the learning from input patterns, a large amount of images with similar image content or tags may result in such a change in values. However, with the learning from massive input patterns, the value of γ_{SA} becomes stable. It demonstrates the convergency of robustness measure.

TABLE 1: Clustering results on NUS-WIDE data set using visual and textual features in terms of nine classes.

Average Precision	dog	bear	cat	bird	flower	lake	sky	sunset	wedding	Overall
K-means	0.8065	0.7691	0.8964	0.6956	0.7765	0.4873	0.5278	0.5836	0.9148	0.7175
CIHC	0.8524	0.8343	0.9167	0.8942	0.8756	0.6544	0.7466	0.6384	0.9127	0.8139
SRC	0.8184	0.7831	0.8193	0.8302	0.8713	0.6852	0.7132	0.5684	0.8723	0.7735
Comrafs	0.8292	0.6884	0.9236	0.8541	0.8667	0.6719	0.7240	0.6562	0.9065	0.7959
NMF	0.8677	0.8133	0.8623	0.7845	0.8259	0.7848	0.7134	0.6956	0.8648	0.8014
SS-NMF	0.8913	0.8272	0.9149	0.8366	0.8723	0.8213	0.7274	0.7346	0.9174	0.8381
Fusion ART	0.8139	0.7914	0.8500	0.9131	0.8368	0.7448	0.7039	0.6829	0.9653	0.8111
GHF-ART	0.9339	0.8814	0.9685	0.9231	0.9368	0.8755	0.7782	0.7829	0.9932	0.8971
GHF-ART(SS)	**0.9681**	**0.9023**	**0.9719**	**0.9655**	**0.9593**	**0.8864**	**0.8132**	**0.8482**	**0.9961**	**0.9234**

6.1.2 Clustering Performance Comparison

We compare the performance of GHF-ART with Fusion ART which is the original model of GHF-ART, the baseline algorithm K-means, and existing co-clustering algorithms CIHC, SRC, Comrafs, NMF and SS-NMF. To make a fair comparison, For K-means, we concatenate the visual and textual features and use Euclidean distance. For K-means, SRC and NMF which need to set the number of clusters and iterations, we average their performance with different cluster numbers ranging from 9 to 15 and set the number of iteration to 50. The parameter settings of Fusion ART are the same with GHF-ART. For fusion ART and SRC which need to set the weights for multi-modal features, we set the value of weight by 0.7 which is the best setting in our empirical study. For the semi-supervised algorithms SS-NMF and GHF-ART(SS), three images of each class are used as user preferences. As CIHC applies ratio cut which only divides the data set into two clusters, we calculate the precision of each class by clustering with each of all other classes and averaging them. As two-class clustering is easier than our nine-class one, the effectiveness of GHF-ART can still be demonstrated if their performance are comparable.

Table 1 shows the clustering performance in average precision for each class using the visual content of images and the corresponding surrounding text. We observe that GHF-ART outperforms the others in all cases. K-means usually achieves the worst result especially for the classes "bird", "lake" and "sky". The reason should be that the sample mean in the concatenated feature space cannot well represent the common characteristics of features for some classes. CIHC, Comrafs and NMF usually achieve comparable performance and outperform SRC. For the semi-supervised algorithms, we can see that SS-NMF and GHF-ART(SS) achieve better performance than their unsupervised version. Besides, GHF-ART outperforms Fusion ART in all classes, which shows the effectiveness of the proposed methods in addressing the limitations of Fusion ART.

To evaluate the scalability of GHF-ART to big data, we study the time cost of each algorithm with the increase in the number of input patterns. Since the user preferences for GHF-ART are given before the clustering, the time cost of GHF-ART(SS) is almost the same as that of GHF-ART. As shown in Fig. 4, along with the increase in the number of patterns, Comrafs has the highest time

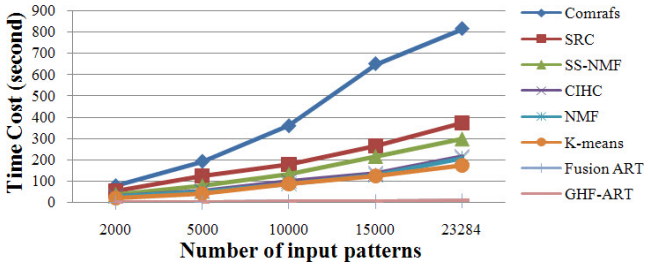

Fig. 4: Time cost of eight algorithms on NUS-WIDE data set along with the increase of input patterns.

cost among all the algorithms. CIHC and NMF have a similar time cost and are slower than K-means. Fusion ART and GHF-ART incur a very small increase of time cost while those of other algorithms increase greatly. Althrough GHF-ART employs the robustness measure, Their time costs are similar. For over 20,000 images, GHF-ART needs less than 10 seconds to complete the clustering process.

To further evaluate the performance of GHF-ART under more complex problems, we run experiments with more classes and noisier data. To this end, we choose nine new classes, including beach, boat, bridge, car, cloud, coral, fish, garden and tree, each of which contains 1500 images. Three classes "car", "cloud" and "tree" are deemed as noisy classes since all the algorithms achieve lower performance. In addition to weighted average precision, we further utilize cluster and class entropies [33], purity [38] and rand index [39] as performance measures. For those algorithms which need a pre-defined number of clusters, we set the number from 18 to 30 and calculate the average performance. For K-means, Fusion ART, GHF-ART and GHF-ART(SS), which are sensitive to initialization, we repeat the experiments for ten times and calculate the means and standard deviations.

Table 2 shows the results on the original data set with 9 classes and the new data set with 18 classes. In Table 2(a), we observe that GHF-ART(SS) achieves the best results in all the evaluation measures in terms of the means. Without supervision, GHF-ART still obtains better performance than all other algorithms. Comparing Table 2(b) with Table 2(a), we find that all algorithms perform worse when the number of classes increases. This is expected as the increase in the number of classes makes it more difficult to partition the feature spaces. However, GHF-ART still obtain the best results.

TABLE 2: Clustering results on NUS-WIDE data set with 9 and 18 classes in terms of weighted average precision, cluster entropy ($H_{cluster}$), class entropy (H_{class}), purity and rand index (RI).

	K-means	CIHC	SRC	Comrafs	NMF	SS-NMF	Fusion ART	GHF-ART	GHF-ART(SS)
Average Precision	0.6582 ± 0.036	0.8139	0.7735	0.7959	0.8014	0.8381	0.8047 ± 0.031	0.8663 ± 0.022	$\mathbf{0.9035 \pm 0.016}$
$H_{cluster}$	0.5317 ± 0.034	0.4105	0.4462	0.4367	0.4189	0.3922	0.4124 ± 0.024	0.3692 ± 0.018	$\mathbf{0.3547 \pm 0.019}$
H_{class}	0.4792 ± 0.037	0.3924	0.4169	0.4386	0.3779	0.3761	0.3744 ± 0.016	0.3583 ± 0.019	$\mathbf{0.3428 \pm 0.013}$
Purity	0.7118 ± 0.029	0.8307	0.7891	0.8036	0.8167	0.8498	0.8352 ± 0.027	0.8863 ± 0.018	$\mathbf{0.9085 \pm 0.021}$
RI	0.6291 ± 0.031	0.7806	0.7485	0.7340	0.7615	0.7759	0.7467 ± 0.018	0.7961 ± 0.023	$\mathbf{0.8216 \pm 0.013}$

(a) Clustering on 9 classes

	K-means	CIHC	SRC	Comrafs	NMF	SS-NMF	Fusion ART	GHF-ART	GHF-ART(SS)
Average Precision	0.4528 ± 0.042	0.7739	0.6812	0.6583	0.7209	0.7637	0.7379 ± 0.024	0.7933 ± 0.023	$\mathbf{0.8366 \pm 0.024}$
$H_{cluster}$	0.6355 ± 0.024	0.4203	0.4726	0.4639	0.4491	0.4215	0.4378 ± 0.024	0.4109 ± 0.018	$\mathbf{0.3921 \pm 0.019}$
H_{class}	0.3892 ± 0.029	0.4161	0.4497	0.4667	0.4018	0.3894	0.4125 ± 0.021	0.3849 ± 0.016	$\mathbf{0.3624 \pm 0.018}$
Purity	0.4682 ± 0.033	0.7795	0.6944	0.6727	0.7279	0.7346	0.7193 ± 0.018	0.8054 ± 0.022	$\mathbf{0.8433 \pm 0.023}$
RI	0.4677 ± 0.028	0.7049	0.6728	0.6496	0.7105	0.7488	0.7245 ± 0.022	0.7523 ± 0.012	$\mathbf{0.7681 \pm 0.014}$

(b) Clustering on 18 classes

TABLE 3: Clustering results on the NUS-WIDE data set using the whole set and the subsets.

		dog	bear	cat	bird	flower	lake	sky	sunset	wedding
Whole Set	Average Precision	0.9339	0.8814	0.9685	0.9231	0.9368	0.8755	0.7782	0.7829	0.9932
	# of clusters	3	2	3	4	2	3	3	1	1
Subsets	Average Precision	0.9273	0.9036	0.9512	0.9039	0.9368	0.8622	0.7694	0.8315	0.9967
	# of clusters	2	2	3	3	2	2	3	2	1

TABLE 4: Clustering results on NUS-WIDE data set by applying equal weights to visual and textual features in all the algorithms. GHF-ART$_{ew}$ indicates using equal weights and GHF-ART$_{aw}$ indicates using adaptive weights

Average Precision	dog	bear	cat	bird	flower	lake	sky	sunset	wedding	Overall
K-means	0.8065	0.7691	0.8964	0.6956	0.7765	0.4873	0.5278	0.5836	0.9148	0.7175
CIHC	0.8524	**0.8343**	0.9167	0.8942	0.8756	0.6544	0.7466	0.6384	0.9127	0.8139
SRC	0.7629	0.7781	0.7667	0.8352	0.8274	0.6903	0.7095	0.5971	0.8566	0.7326
Comrafs	0.8292	0.6884	**0.9236**	0.8541	0.8667	0.6719	0.7240	0.6562	0.9065	0.7959
NMF	0.8677	0.8133	0.8623	0.7845	0.8259	0.7848	0.7134	0.6956	0.8648	0.8014
Fusion ART	0.7960	0.7835	0.8376	0.8891	0.8267	0.7614	0.6850	0.7035	0.9661	0.8037
GHF-ART$_{ew}$	**0.8746**	0.7812	0.9211	**0.9046**	**0.8952**	**0.8748**	**0.7814**	**0.7585**	**0.9746**	**0.8629**
GHF-ART$_{aw}$	0.9339	0.8814	0.9685	0.9231	0.9368	0.8755	0.7782	0.7829	0.9932	0.8971

To evaluate statistical significance of performance difference, we conduct t-test among Fusion ART, GHF-ART and GHF-ART(SS). The results show that the performance of Fusion ART and GHF-ART is significantly different at 0.05 level of significance in all evaluation measures except class entropy, of which the difference is at 0.1 level. For GHF-ART and GHF-ART(SS), the difference between their performance in average precision, purity and rand index is significant at 0.05 level of significance. For cluster entropy and class entropy, the performance difference is at 0.1 level.

6.1.3 Evaluation on Incremental Property

To evaluate the incremental property of GHF-ART, as described in Section 5, we divide the original data set with nine classes into four smaller subsets and apply GHF-ART to them sequentially. Then, we compare the clustering performance of GHF-ART with that for the whole data set. To make a fair comparison, we randomize the sequence of input patterns in all the subsets.

As are shown in Table 3, we observe that, for all the classes, the number of clusters and average precision are similar for clustering the whole data set and the subsets. This shows that, given several sequential data sets with random patten sequences, the cluster structure obtained by clustering the whole data set and the subsets are similar. This demonstrates that GHF-ART is able to cluster the new patterns of the updated data set by incrementally adapting the cluster structure learnt from the original data set.

6.1.4 Case Study Analysis of Performance

We present a case study to analyze why GHF-ART outperforms other algorithms. Since one major difference between GHF-ART and the other algorithms is the adaptive weighting method of GHF-ART, we evaluate the performance when all the algorithms employ equal weights for the visual and textual features. The results are summarized in Table 4. The performance of GHF-ART with adaptive weights (GHF-ART$_{aw}$) is also listed at below for a comparison. Comparing with GHF-ART$_{aw}$, the performance of GHF-ART with equal weights (GHF-ART$_{ew}$) has an obvious decrease in most classes, especially for the class "bear". Similarly, the performance of Fusion ART and SRC also have a decrease when using the equal weights. It demonstrates the importance of weighting for feature modalities in clustering. However, GHF-ART$_{ew}$ still obtains the best results in six out of nine classes.

In addition, suppose we use the learning function of

Fuzzy ART instead of the proposed learning method for the meta-information, GHF-ART degenerates to the original Fusion ART. We see that Fusion ART achieves comparable performance with NMF and a little bit lower than CIHC in the overall performance. For specific classes, Fusion ART obtains the best result in "wedding" and usually achieves a comparable performance for the other classes. However, with our proposed meta-information learning method, GHF-ART$_{ew}$ outperforms Fusion ART in most classes and has a relatively big improvement in "lake", "sky" and "sunset". This also demonstrates that the proposed learning method of meta-information enables GHF-ART to be robust in handling noisy text.

In comparison, we find all the other algorithms achieve a low level of performance on these noisy classes. This, we reckon, is due to the differences between various methods in handling the patterns. For example, K-means generates hyperspherical clusters in the feature space which are sensitive to noise. Therefore, K-means performs poorly in the noisy classes but obtains comparable performance in classes such as "wedding". CIHC and SRC, which employ spectral clustering, derive eigenvectors from the graph affinity matrices. As such, the noisy features may lead to spurious correlations between patterns. This is why CIHC obtains reasonable performance in all the classes except the three noisy classes. Since SRC employs K-means to get the final clusters, it also suffers from the drawbacks of K-means. NMF derives the cluster indicator matrix from the relational matrices which maps the data into a non-negative latent semantic space. Similar to spectral clustering, noisy features should also be the main reason for the poor performance in the noisy classes. Comrafs performs clustering by finding a cluster structure of patterns that maximizes the Most Probable Explanation based on mutual information. Therefore, noisy features affect the calculation of mutual information and lead to incorrect classification of patterns.

Based on the above analysis, we may conclude that GHF-ART outperforms the other algorithms when the surrounding text is noisy and when the desired weights for different feature modalities are not equal.

6.2 Corel Data Set

Corel data set is a subset of Corel CDs data set and consists of 5,000 images from 50 Corel Stock Photo CDs, each of which contains 100 images on the same topic. Each image is annotated by an average of 3-5 keywords from a dictionary of 374 words. We utilize the images of six classes including "sunset", "plane","birds","bear","beach" and "hills". Similar to the NUS-WIDE data set, we extract the 426 visual features and build the textual features using 374 words.

6.2.1 Performance of Robustness Measure

Similar to the NUS-WIDE data set, we test the performance of GHF-ART with different settings of contribution parameter of textual features on Corel data set. In

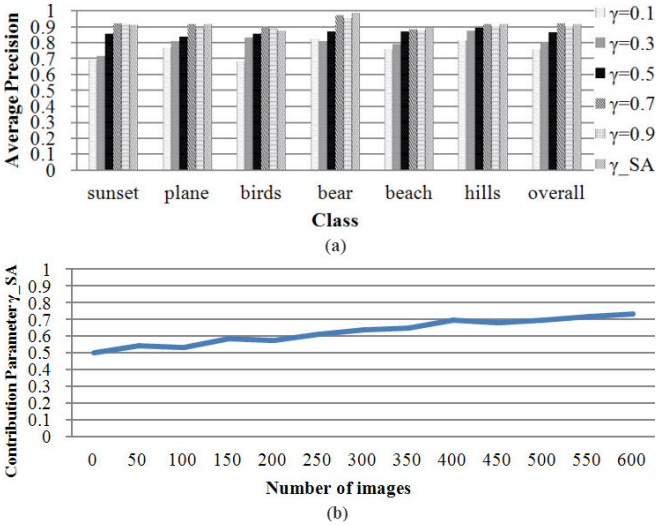

Fig. 5: (a) Clustering performance using fixed contribution parameters (γ) and self-adapted contribution parameter (γ_{SA}); (b) Tracking of γ_{SA} on Corel data set.

Fig. 5(a), we observe that robustness measure achieves the best results for most classes except "sunset" and "birds" and the best overall performance is achieved by $\gamma = 0.7$. However, it still outperforms the other settings and achieves performance very close to the best setting. The value tracking of γ is shown in Fig. 5(b). In contrast to that for NUS-WIDE, the result shows a relatively smooth change in the contribution parameter value. The reason should be that the Corel data set contains less noisy tags. We can see that the value gradually increases and stabilizes at $\gamma = 0.7$. It demonstrates that the robustness measure can effectively adjust the contribution parameter to the best setting.

6.2.2 Clustering Performance Comparison

Similar to the NUS-WIDE data set, we evaluate the performance of GHF-ART in terms of average precision, cluster and class entropies, purity and rand index. We set the number of clusters ranging from 6 to 15 for those algorithms which need a pre-defined number of clusters. As shown in Table 5, firstly, we observe that all algorithms achieve better clustering performance than that of NUS-WIDE data set. One possible reason is that the visual content of the images belonging to the same category is more similar and the tags of Corel data set is relatively cleaner. We can also see that GHF-ART and GHF-ART(SS) outperform the other algorithms in all the performance measures. Particularly, GHF-ART got a close mean result to CIHC and SS-NMF in average precision, class entropy, purity and Rand Index but a much better performance in cluster entropy. With supervisory information, GHF-ART(SS) has a further improvement on GHF-ART. In addition, GHF-ART has a big improvement on Fusion ART, which demonstrates the effectiveness of our proposed adaptive feature weighting and meta-information learning methods in improving the performance and robustness of Fusion ART.

TABLE 5: Clustering results on Corel data set using visual content and surrounding text.

	K-means	CIHC	SRC	Comrafs	NMF	SS-NMF	Fusion ART	GHF-ART	GHF-ART(SS)
Average Precision	0.7245 ± 0.023	0.8940	0.8697	0.8115	0.8794	0.8960	0.8525 ± 0.027	0.8944 ± 0.018	$\mathbf{0.9168 \pm 0.019}$
$H_{cluster}$	0.3816 ± 0.024	0.2614	0.2803	0.3316	0.2771	0.2592	0.2409 ± 0.019	0.2184 ± 0.016	$\mathbf{0.1960 \pm 0.014}$
H_{class}	0.3538 ± 0.025	0.2566	0.2714	0.2972	0.2703	0.2667	0.2793 ± 0.022	0.2521 ± 0.018	$\mathbf{0.2366 \pm 0.015}$
Purity	0.7263 ± 0.026	0.9031	0.8725	0.8304	0.8862	0.8997	0.8628 ± 0.023	0.8975 ± 0.021	$\mathbf{0.9176 \pm 0.015}$
RI	0.6635 ± 0.024	0.8347	0.8051	0.7734	0.8172	0.8416	0.8116 ± 0.015	0.8342 ± 0.018	$\mathbf{0.8533 \pm 0.014}$

TABLE 6: Clustering results on Corel data set using visual content, surrounding text and category information.

	K-means	CIHC	SRC	Comrafs	NMF	SS-NMF	Fusion ART	GHF-ART	GHF-ART(SS)
Average Precision	0.7254 ± 0.020	0.9014	0.8782	0.8279	0.8865	0.9047	**1**	**1**	**1**
$H_{cluster}$	0.3688 ± 0.022	0.2544	0.2758	0.3263	0.2709	0.2537	0.1727 ± 0.023	0.1496 ± 0.016	$\mathbf{0.1362 \pm 0.014}$
H_{class}	0.3251 ± 0.026	0.2467	0.2682	0.2543	0.2489	0.2466	**0**	**0**	**0**
Purity	0.7284 ± 0.020	0.9106	0.8721	0.8463	0.8917	0.9044	**1**	**1**	**1**
RI	0.6775 ± 0.021	0.8428	0.8147	0.8045	0.8276	0.8315	0.9061 ± 0.019	0.9297 ± 0.021	$\mathbf{0.9485 \pm 0.016}$

Similar to NUS-WIDE data set, we further conduct t-test between the performance of Fusion ART, GHF-ART and GHF-ART(SS) reported in Table 5. The results show that the performance differences between Fusion ART, GHF-ART and GHF-ART(SS) are significant at 0.05 level of significance across all evaluation measures.

6.2.3 Clustering Performance Comparison with Category Information

We further conduct the experiments by incorporating the category information for clustering. The category information is used in the same way of surrounding text. In view that the category information for each image is exactly one word, it therefore can also be seen as a annotation corpus without any noise. As shown in Table 6, we observe that Fusion ART, GHF-ART and its semi-supervised version GHF-ART(SS) achieve 100% in average precision, class entropy and purity. It is because the ART-based algorithms not only have a global optimization between features but also constraints for each feature modality. Therefore, with the category label, the ART-based algorithms can effectively identify the classes of images. Besides, we can also observe an improvement of GHF- ART(SS) over Fusion ART and GHF-ART in cluster entropy and rand index, which also consider how the patterns with the same label are grouped together.

Comparing the results with those in Table 5, we can find that Fusion ART, GHF-ART and GHF-ART(SS) also obtain a big improvement in terms of cluster entropy and rand index, while the other algorithms have a relatively small improvement. The reason should be that the global optimization considers the overall similarity across all the feature channels so that the noisy features still contribute to incorrect categorization. It demonstrates the importance of taking in the fitness of patterns in terms of the overall similarity and also that for different modalities individually rather than the only global optimum.

6.3 20 Newsgroups Data Set

The 20 Newsgroups data set [9] is a popular public data set which comprises nearly 20,000 newsgroup documents across 20 different newsgroups and is widely used for the experiments of text clustering techniques.

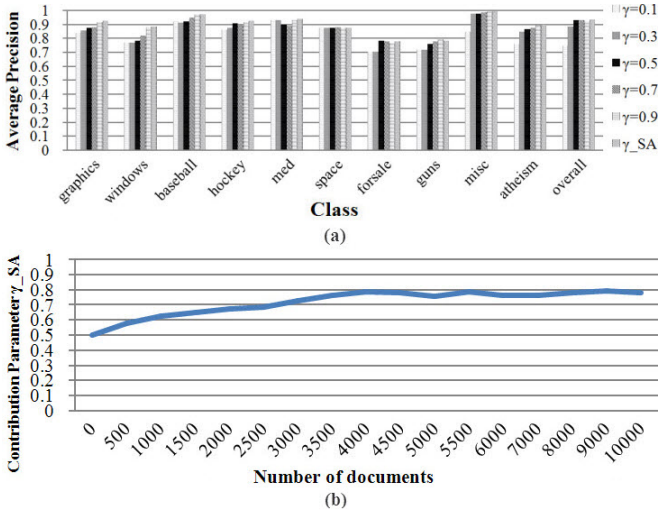

Fig. 6: (a) Clustering performance using fixed contribution parameters (γ) and self-adapted contribution parameter (γ_{SA}); (b) Tracking of γ_{SA} on 20 newsgroups data set.

We directly collect ten classes from the processed matlab version of the 20news-bydate data set and each of them contains nearly 1,000 documents. For the ease of discussion, we refer the ten categories by the abbreviations as follows: comp.graphics (graphics), comp.windows.x (windows), rec.sport.baseball (baseball), rec.sport.hockey (hockey), sci.med (med), sci.space (space), misc.forsale (forsale), talk.politics.guns (guns), talk.politics.misc (misc) and alt.atheism (atheism). We use the traditional text mining algorithm tf-idf to extract the features of documents and use the words in the category information to construct the category features.

6.3.1 Performance of Robustness Measure

Fig. 6 shows the clustering results with different settings of contribution parameter of the category features. In Fig. 6(a), we observe that the robustness measure works well for all classes and usually produces the best results. From Fig. 6(b), we can observe that the contribution parameter of category features gradually increases from 0.5 to over 0.6 after 1,500 input patterns. Despite the small fluctuation, the value stabilizes at around 0.8,

TABLE 7: Clustering results on 20 Newsgroups data set using document content and category information.

	K-means	CIHC	SRC	Comrafs	NMF	SS-NMF	Fusion ART	GHF-ART	GHF-ART(SS)
Average Precision	0.6386 ± 0.027	0.7583	0.7246	0.6547	0.7357	0.7869	0.7566 ± 0.021	0.8071 ± 0.023	$\mathbf{0.8452 \pm 0.018}$
$H_{cluster}$	0.5284 ± 0.031	0.4573	0.4630	0.5162	0.4487	0.4296	0.4469 ± 0.015	0.4131 ± 0.017	$\mathbf{0.3824 \pm 0.019}$
H_{class}	0.4833 ± 0.025	0.4246	0.4432	0.4679	0.4267	0.3938	0.4016 ± 0.016	0.3822 ± 0.018	$\mathbf{0.3642 \pm 0.018}$
Purity	0.6826 ± 0.027	0.7711	0.7348	0.6950	0.7503	0.7836	0.7538 ± 0.021	0.7994 ± 0.018	$\mathbf{0.8435 \pm 0.021}$
RI	0.6670 ± 0.025	0.7284	0.6867	0.6136	0.7019	0.7458	0.7268 ± 0.017	0.7759 ± 0.022	$\mathbf{0.8013 \pm 0.019}$

which indicates that the category information is more robust during the clustering process.

6.3.2 Clustering Performance Comparison

Similar to the NUS-WIDE data set, we evaluate the clustering performance of GHF-ART using average precision, cluster and class entropies, purity and rand index. Since the number of classes in 20 Newsgroups data set is 10, we set the number of clusters ranging from 10 to 15. In Table 7, we can see that GHF-ART and GHF-ART(SS) outperform the other algorithms in all the performance measures. Moreover, both of them achieve higher than 80% in average precision and purity while the other algorithms typically obtain less than 75% except CIHC and SS-NMF. Similarly, a gain of more than 3% over the best performance by the other algorithms is achieved in rand index. The t-test results further show that the performance of Fusion ART, GHF-ART and GHF-ART(SS) are significantly different at 0.05 level of significance in all evaluation measures. In fact, we observe that GHF-ART has a big improvement over Fusion ART. This demonstrates that the proposed feature weighting algorithm and meta-information learning method can help to improve the performance of Fusion ART in the heterogeneous data co-clustering task.

7 CONCLUSIONS

In this paper, we have proposed a novel heterogeneous co-clustering algorithm termed Generalized Heterogeneous Fusion ART (GHF-ART) aiming at fast and robust clustering of web multimedia data. GHF-ART extends the Heterogeneous Fusion ART from two channels to multiple channels so that GHF-ART can be applied to the clustering of more than two modalities wherein each channel may receive different types of data patterns. By generalizing the feature construction methods for multimedia documents and incorporating an adaptive channel weighting algorithm, GHF-ART is able to effectively integrate different types of features across multiple pattern channels for measuring pattern similarity.

Comparing with existing co-clustering algorithms [2], [3], [4], [5], GHF-ART has the advantages in four aspects: 1) **Strong noise immunity**: GHF-ART models the textual features of meta-information by the probability distribution of tag occurrences so that the key tags of clusters can be incrementally identified while the noisy tags are depressed. This helps to maintain the robustness of GHF-ART when the quality of text is low; 2) **Adaptive channel weighting method**: GHF-ART has a well-defined weighting algorithm for multi-modal feature channels. Different from the modality selection method in SS-NMF [4] which only learn the weights from the prior knowledge in the distance learning step, GHF-ART evaluates the weights of feature modalities by incrementally learning from the intra-cluster scatters of so that the importance of feature modalities in clustering can be incrementally evaluated, which increases the robustness of GHF-ART in fusing feature modalities for measuring pattern similarity. 3) **Low computational complexity**: The real-time cluster searching mechanism of GHF-ART leads to a linear time complexity, analyzed in Section 5.6, which enables GHF-ART to be scalable to big data sets; and 4) **Incremental clustering manner**: Web multimedia data is usually big and requires frequent update. As mentioned in Section 2, existing co-clustering methods typically make use of an global objective functions, which is then solved by an iterative optimization approach. When new data are available, these methods will have to be re-run on the entire data set. In contrast, GHF-ART can re-cluster the new data set by adapting the original cluster structure incrementally. This so-called incremental property of GHF-ART is theoretically guaranteed by the ART clustering mechanism, which incrementally clusters the input patterns, one at a time, into the clustering structure. In this way, GHF-ART is able to cluster the new patterns in an incremental manner without referring to the old data by adapting existing clusters or creating a cluster when a new pattern is distinct from existing clusters.

Going forward, there remain some issues for further investigation. Firstly, tag ranking methods can be employed in the textual feature construction stage to filter noisy tags or give more weight to key tags so as to further depress the effect of noisy tags. Secondly, since the learning function for meta-information is designed to track the probabilistic distribution of the data set in an incremental manner, there is no guarantee of convergence in response to the changing data characteristics. Thirdly, as the current method for tuning vigilance parameters still cannot fully solve the problem of category proliferation, developing effective criteria for learning the desired vigilance parameters values will also be in our future work.

ACKNOWLEDGMENT

This study is supported by the Singapore National Research Foundation under its Interactive & Digital Media (IDM) Public Sector R&D Funding Initiative (Grant No. NRF2008IDMIDM004-018) administered by the IDM Programme Office.

REFERENCES

[1] B. Gao, T.-Y. Liu, X. Zheng, Q.-S. Cheng, and W.-Y. Ma, "Consistent bipartite graph co-partitioning for star-structured high-order heterogeneous data co-clustering," *Proc. of Int'l Conference on Knowledge Discovery and Data Mining*, pp. 41–50, 2005.

[2] M. Rege, M. Dong, and J. Hua, "Graph theoretical framework for simultaneously integrating visual and textual features for efficient web image clustering," *Proc. of Int'l Conference on World Wide Web*, pp. 317–326, 2008.

[3] B. Long, X. Wu, Z. Zhang, and P. Yu, "Spectral clustering for multi-type relational data," *In ICML*, pp. 585–592, 2006.

[4] Y. Chen, L. Wang, and M. Dong, "Non-negative matrix factorization for semisupervised heterogeneous data coclustering," *In TKDE*, pp. 1459–1474, 2010.

[5] R. Bekkerman and J. Jeon, "Multi-modal clustering for multimedia collections," *In CVPR*, pp. 1–8, 2007.

[6] X. Hu, N. Sun, C. Zhang, , and T.-S. Chua, "Exploiting internal and external semantics for the clustering of short texts using world knowledge," *Proc.of ACM conference on Information and knowledge management*, pp. 919–928, 2009.

[7] L. Meng and A.-H. Tan, "Heterogeneous learning of visual and textual features for social web image co-clustering," *Technical Report*, 2012.

[8] T. 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 CIVR*, pp. 1–9, 2009.

[9] K. Lang, "Newsweeder: Learning to filter netnews," *Proc. Int'l Conf. Machine Learning*, pp. 331–339, 2005.

[10] S. Harabagiu and F. Lacatusu, "Using topic themes for multi-document summarization," *ACM Transactions on Information Systems*, vol. 28, no. 3, pp. 1–47, 2010.

[11] D. Wang, S. Zhu, T. Li, Y. Chi, and Y. Gong, "Integrating document clustering and multidocument summarization," *ACM Transactions on Knowledge Discovery from Data*, vol. 5, no. 3, pp. 1–26, 2011.

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

[13] A. Messina and M. Montagnuolo, "A generalised cross-modal clustering method applied to multimedia news semantic indexing and retrieval," *In WWW*, pp. 321–330, 2009.

[14] N. Rasiwasia and J. Pereira, "A new approach to cross-modal multimedia retrieval," *In MM*, pp. 251–260, 2010.

[15] M. Li, X.-B. Xue, and Z.-H. Zhou, "Exploiting multi-modal interactions: A unified framework," *In IJCAI*, pp. 1120–1125, 2009.

[16] R. Zhao and W. Grosky, "Narrowing the semantic gap improved text-based web document retrieval using visual features," *IEEE Transactions on Multimedia*, pp. 189–200, 2002.

[17] D. Cai, X. He, Z. Li, W. Ma, and J. Wen, "Hierarchical clustering of www image search results using visual, textual and link information," *Proc. of ACM Multimedia*, pp. 952–959, 2004.

[18] T. Jiang and A.-H. Tan, "Learning image-text associations," *IEEE Transactions on Knowledge and Data Engineering*, pp. 161–177, 2009.

[19] ——, "Discovering image-text associations for cross-media web information fusion," *In PKDD*, pp. 561–568, 2006.

[20] A.-H. Tan, G. A. Carpenter, and S. Grossberg, "Intelligence through interaction: Towards a unified theory for learning," *In LNCS*, vol. 4491, pp. 1094–1103, 2007.

[21] B. Gao, T. Liu, T. Qin, X. Zheng, Q. Cheng, and W. Ma, "Web image clustering by consistent utilization of visual features and surrounding texts," *Proc. of ACM Multimedia*, pp. 112–121, 2005.

[22] M. Renge, M. Dong, , and F. Fotouhi, "Co-clustering documents and words using bipartite isoperimetric graph partitioning," *Proc. of Int'l Conference on Data Mining*, pp. 532–541, 2006.

[23] X. Cai, F. Nie, H. Huang, and F. Kamangar, "Heterogeneous image feature integration via multi-modal spectral clustering," *In CVPR*, pp. 1977–1984, 2011.

[24] Z. Fu, H. H. S. Ip, H. Lu, and Z. Lu, "Multi-modal constraint propogation for heterogeneous image clustering," *In MM*, pp. 143–152, 2011.

[25] W. Xu, X. Liu, and Y. Gong, "Document clustering based on non-negative matrix factorization," *Proc. of SIGIR conference on Research and development in informaion retrieval*, pp. 268–273, 2003.

[26] Q. Gu and J. Zhou, "Co-clustering on manifolds," *In KDD*, pp. 359–367, 2009.

[27] R. Bekkerman, M. Sahami, and E. Learned-Miller, "Combinatorial markov random fields," *In ECML*, pp. 30–41, 2006.

[28] R. Bekkerman and M. Sahami, "Semi-supervised clustering using combinatorial mrfs," *In ICML Workshop on Learning in Structured Output Spaces*, 2006.

[29] R. Bekkerman, M. Scholz, and K. Viswanathan, "Improving clustering stability with combinatorial mrfs," *In KDD*, pp. 99–108, 2009.

[30] D. Liu, X. Hua, L. Yang, M. Wang, and H. Zhang, "Tag ranking," *Proc. of Int'l Conference on World Wide Web*, pp. 351–360, 2009.

[31] X. Li, C. G. M. Snoek, and M. Worring, "Tag relevance by neighbor voting for social image retrieval," *Proc. of ACM Multimedia*, 2008.

[32] G. Carpenter and S. Grossberg, "A massively parallel architecture for a self-organizing neural pattern recognition machine," *Computer Vision, Graphics, and Image Processing*, pp. 54–115, 1987.

[33] A.-H. Tan, H.-L. Ong, H. Pan, J. Ng, and Q. Li, "Towards personalized web intelligence," *Knowledge and Information Systems*, pp. 595–616, 2004.

[34] L. Nguyen, K. Woon, and A.-H. Tan, "A self-organizing neural model for multimedia information fusion," *International Conference on Information Fusion*, pp. 1–7, 2008.

[35] G. A. Carpenter, S. Grossberg, and D. B. Rosen, "Fuzzy ART: Fast stable learning and categorization of analog patterns by an adaptive resonance system," *Neural Networks*, pp. 759–771, 1991.

[36] G. A. Carpenter, S. Grossberg, and J. Reynolds, "ARTMAP: Supervised real-time learning and classification of nonstationary data by a self-organizing neural network," *Neural Networks*, pp. 565–588, 1991.

[37] A.-H. Tan, "Adaptive resonance associative map," *Neural Networks*, pp. 437–446, 1995.

[38] Y. Zhao and G. Karypis, "Criterion functions for document clustering: experiments and analysis," *Technical Report*, 2001.

[39] R. Xu and D. C. W. II, "Bartmap: A viable structure for biclustering," *Neural Networks*, pp. 709–716, 2011.

Lei Meng received the B.Eng. degree in the Department of Computer Science and Technology from Shandong University, China, in 2010. He is currently a Ph.D. student in the School of Computer Engineering, Nanyang Technological University, Singapore. His research interests include machine learning, heterogeneous data co-clustering and web media data mining.

Ah-Hwee Tan received a PhD in cognitive and neural systems from Boston University, a Bachelor of Science (First Class Honors) and a Master of Science in computer and information science from the National University of Singapore. He is currently an associate professor and the head of Division of Software and Information Systems at the School of Computer Engineering (SCE), Nanyang Technological University. Prior to joining NTU, he was a research manager at the A*STAR Institute for Infocomm Research (I^2R), spearheading the Text Mining and Intelligent Agents research programmes. His current research interests include brain-inspired intelligent agents, cognitive and neural systems, machine learning, knowledge discovery and text mining.

Dong Xu is currently an associate professor at Nanyang Technological University (NTU) in Singapore. He received the B.Eng. and PhD degrees from University of Science and Technology of China, in 2001 and 2005, respectively. During his PhD study, he worked at Microsoft Research Asia and The Chinese University of Hong Kong for more than two years. He also worked at Columbia University for one year as a postdoctoral research scientist. His research focuses on new theories, algorithms and systems for intelligent processing and understanding of visual data such as images and videos.