Semi-supervised hierarchical clustering for personalized web image organization

Lei MENG

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

Citation

MENG, Lei and TAN, Ah-hwee. Semi-supervised hierarchical clustering for personalized web image organization. (2012). *Proceedings of International Joint Conference on Neural Networks (IJCNN), Australia, June 10-15*.

Available at: https://ink.library.smu.edu.sg/sis_research/6887

Semi-supervised hierarchical clustering for personalized web image organization

Conference Paper · June 2012

DOI: 10.1109/IJCNN.2012.6252397

Lei Meng
National University of Singapore

Ah-Hwee Tan
Singapore Mamagement University

Semi-supervised Hierarchical Clustering for Personalized Web Image Organization

Lei Meng
School of Computer Engineering
Nanyang Technological University
Singapore
meng0027@ntu.edu.sg

Ah-Hwee Tan
School of Computer Engineering
Nanyang Technological University
Singapore
asahtan@ntu.edu.sg

***Abstract*—Existing efforts on web image organization usually transform the task into surrounding text clustering. However, Current text clustering algorithms do not address the problem of insufficient statistical information for image representation and noisy tags which greatly decreases the clustering performance while increases the computational cost. In this paper, we propose a two-step semi-supervised hierarchical clustering algorithm, Personalized Hierarchical Theme-based Clustering (PHTC), for web image organization. In the first step, the Probabilistic Fusion ART (PF-ART) is proposed for grouping semantically similar images and simultaneously learning the probabilistic distribution of tag occurrence for mining the key tags/topics of clusters. In this way, the side-effect of noisy tags can be largely eliminated. Moreover, PF-ART can incorporate user preference for semi-supervised learning and provide users a direct control of clustering results. In the second step, a novel agglomerative merging strategy based on Cluster Semantic Relevance, proposed for measuring the semantic similarity between clusters, is employed for associating the clusters by generating a semantic hierarchy. Different from existing hierarchical clustering algorithms, the proposed merging strategy can provide a multi-branch tree structure which is more systematic and clearer than traditional binary tree structure. Extensive experiments on two real world web image data sets, namely NUS-WIDE and Flickr, demonstrate the effectiveness of our algorithm for large web image data sets.**

I. Introduction

Along with the explosive popularity of social web sites, a massive number of web images has appeared in diverse content online. It leads to the need for effective image organization to make the information more systematic and manageable. Two research challenges have been identified. The first challenge is how to learn the semantics (i.e. themes/topics) from images. Most of the existing applications [1]–[4] are based on text clustering techniques, in which the tags of images extracted from their surrounding text (titles, categories information and user descriptions etc.) are used for image representation. This is because current state-of-the-art visual feature extraction techniques cannot fully represent the image content at the semantic level, a problem known as semantic gap. Thus, the problem of image organization is usually transformed into short text categorization. However, similar to the short document categorization problem [5], the tags cannot provide sufficient statistic information for effective similarity measure, i.e. the key tags that are useful for image topic representation cannot be revealed by traditional word weighting strategies, like term frequency-inverse document frequency (tf-idf). Besides, as users usually give descriptions based on their own views, the tags for images in one topic may be diverse, which is known as the problem of noisy tags. Therefore, traditional text clustering algorithms [6]–[8] may fail to achieve reasonable results when they are directly applied on this task. Besides, as existing algorithms are based on computational models, the noisy tags will significantly increase their computational cost.

The second challenge is how to associate the discovered topics. For a real world web image collection, there should be a large number of topics and sub-topics. Some of them may be relevant (e.g. ”white tiger” and ”Indian tiger”), and some of them may belong to a more general topic (e.g. ”tiger” and ”bird” belong to ”animal”). It may result in the generation of too many categories. Therefore, a semantic hierarchy that can reveal the relationship between topics is necessary. However, Existing hierarchical clustering approaches like [4] follow the agglomerative strategy which merges two clusters in one round. It leads to the problem that the generated binary tree structure becomes too complex when the number of the generated clusters is large.

In this paper, we present a two-step hierarchical clustering algorithm termed Personalized Hierarchical Theme-based Clustering (PHTC) for large-scale web image collection organization. PHTC can incrementally discover the semantic categories and the key themes according to user preferences at the same time, and further organize the generated clusters into a multi-branch tree hierarchy. In the first step, we propose a novel semi-supervised clustering algorithm called Probabilistic Fusion Adaptive Resonance Theory (PF-ART), a variant of Fusion ART [9], for generating semantic clusters according to user preferences. Different from Fusion ART, PF-ART represents each cluster using the probabilistic distribution of tag occurrence. Beyond existing semi-supervised clustering algorithms [10], [11], PF-ART not only incorporates relevance feedback to enhance the clustering quality, but also provides the flexibility for users to directly control the degree of topic mining. That is, users can decide whether the clusters are generated according to general topics like ”lion” and ”bird”, or more specific topics like ”lion in zoo” and ”lion in Africa”. In the second step, we propose a similarity measure between categories called Cluster Semantic Relevance (CSR) and an agglomerative merging strategy based on CSR for generating the semantic hierarchy. Different from typical agglomerative algorithms [4], [12]–[14], the proposed algorithm can recognize if the relationship between selected clusters is father and child according to the inner CSR of children categories of the given category. Therefore, the generated hierarchy provides a multi-branch tree structure which is more systematic and clear. We evaluate the performance of our framework on two data sets, namely the NUS-WIDE and Flickr data sets. The experimental results compared with related methods, in terms of the clustering quality as well as time cost, demonstrate the effectiveness of the proposed algorithm.

The rest of paper is organized as follows. Section 2 provides a review of related text clustering techniques and semi-supervised clustering. Section 3 presents the problem statement and proposed approach. The details of our proposed methods are presented in section 4 and section 5. In section 6, extensive experiments on two real world web image data sets are presented to evaluate the performance of our framework. The last section summarizes our work.

II. Related Work

Our work is related to text clustering. Existing methods are usually based on neural networks [3], graph theory [2], [6], Matrix Factorization [7], [8] and hierarchical clustering [2]–[4]. The method used in [3] is a variant of K-means clustering algorithm, which iteratively updates the pre-assigned k cluster centers to minimize the global within-cluster distance cost. The graph based clustering algorithms usually construct an affinity matrix which represents the similarity between vertices and tries to obtain the optimal cut which minimize the cost of edges across two parts. The method used in [2], a variant of Normalized Cut, models the similarity between documents using word frequency and can be solved by eigenvalue decomposition. Differently, the Isoperimetric Co-clustering Algorithm [6] models the relations between documents and words and can get the final cut by solving a sparse system of linear equations. Non-Negative Matrix Factorization [7] can decompose the document-word matrix into two matrices with minimum reconstruction lost. This method can derive a latent semantic space which reveals the relations between each document and a pre-defined number of topics (the axis). The cluster membership of each document is determined by the largest projection value among all topics. Existing hierarchical clustering algorithms for image organization follow two ideas. One is traditional agglomerative hierarchical clustering such as [4]. It starts with each image as a leaf cluster and merges pairs of the most similar clusters in each round. Another approach [2], [3] follow the idea of generating semantic categories using textual features and then grouping visually similar images under each category as the second layer.

Our work is also related to semi-supervised clustering. Existing works [8], [10], [11] are based on relevance feedback where a user indicates if two images belong to one category or not. These feedbacks are incorporated into their objective functions as penalties to improve the clustering results. Different from existing methods, PF-ART receives the user feedback by generating pre-defined categories. These categories will encode similar patterns during the following clustering process and be presented to the users within final clusters. Therefore, users have more chance to get the interesting results. Besides, PF-ART provides a controller called vigilance parameter such that users can have a direct control of the clustering results.

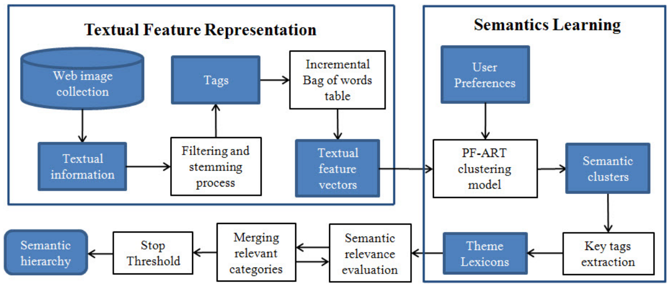

Fig. 1. Procedures of the proposed clustering framework.

III. Problem Statement and Approach

We define the problem of discovering semantics from web images as the problem of mining key themes from the surrounding text of web images. Consider a set of images and the corresponding raw text in the textual description of the original web page, including the title, categories and user descriptions. As the raw text information obtained from web pages is typically noisy, the first task is to filter noisy words, including stop-words, typos and chronic slangs. However, after removing the noisy words and stemming the variation form of words to the root form, the remaining tags are usually diverse because of the diverse views of users. It leads to the difficulties of identifying the potential topics of images. From the perspective that semantically related images hold similar textual description, we apply clustering algorithm for grouping images and mining the key tags of each category. This work can also be treated as a tag refining procedure. To provide a systematic view of clusters, we further associate the semantic clusters by a semantic hierarchy.

The proposed clustering framework (Fig.1) comprises three main modules: 1) Textual feature representation; 2) Semantic categories and theme lexicon generation; and 3) Semantic hierarchy construction. In the first module, given a collection of web images, the associated textual information goes through a pre-processing step so as to obtain salient tags that are meaningful for representing the semantics of the respective images. Then the bag-of-words method is applied for acquiring textual features that represent the presence of tags of each image. Subsequently, in the second module, PF-ART categorizes the images and simultaneously learns the probabilistic distribution of tag occurrence of each category. In addition, user preferences can be incorporated for improving the clustering quality and the degree of topic mining in the final results. The probability distribution is then used to identify the potential key tags (i.e. themes) which constitute the theme lexicons for the respective categories. In the last module, cluster semantic relevance (CSR) is used for evaluating the semantic relevance

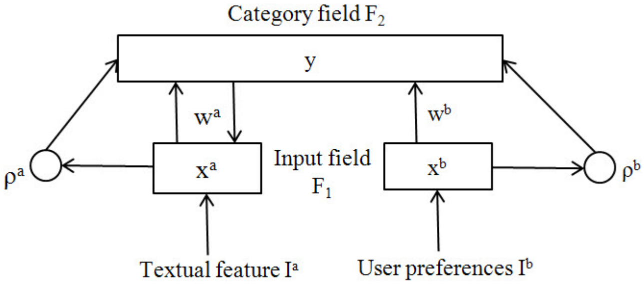

Fig. 2. The Architecture of Probabilistic Fusion ART.

between categories and their children categories such that the merging strategy may determine if two categories can be grouped into a new category or one category should be a child category of the other one. When the highest semantic relevance score of categories reaches the stop threshold, we can obtain a semantic hierarchy where the relationship between categories is revealed by a multi-branch tree structure and the themes of father categories are more general than their children. The details of PF-ART and agglomerative merging strategy are described in the following sections.

IV. Probabilistic Fusion ART

Adaptive Resonance Theory (ART) [15] is a neural theory on how brain processes information. It models clusters as memory prototypes and encodes each input pattern incrementally by a two-way similarity measure. 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 incremental measure. For document clustering domain [16], ART shows a strong noise immunity. Fusion ART [9] extends Fuzzy ART from single input field to multiple ones and provides a general architecture for simultaneously learning of multi-modal feature mappings.

Probabilistic Fusion ART, a variant of Fusion ART, employs a two-channel Fusion ART model (Fig.2), where feature vector I^a encodes the textual feature and preference vector I^b encodes the user-provided tags. Note that the preference vector is only used for incorporating user preferences. x^a and x^b are the received input vectors respectively. w^a and w^b are the weight vector of the respective prototypes of the clusters in the category field. Different from Fusion ART, PF-ART models the cluster prototypes/weights by the probabilistic distribution of tag occurrence, because the original learning strategy cannot preserve sub-topics. The details of PF-ART are described in the following sub-sections.

A. Textual Feature Representation

We construct the textual feature vector based on a textual table consisting of all distinct tags in the whole image set expressed by $\mathbf{t} = [t_1, \ldots, t_m]$. We denote the textual feature for the k^{th} image as $\mathbf{t}^k = [t_1^k, \ldots, t_m^k]$, where t_m^k corresponds to the m^{th} tag in T.Assuming the tag list of img_k is φ_k, the value of t_m^k is given by the impulse response, defined by:

$$\delta(t_m^k, img_k) = \begin{cases} 1 & t_m \in \varphi_k \\ 0 & others \end{cases}.$$

We do not follow traditional methods like tf-idf to weigh each word, because the extracted tags cannot provide sufficient statistical information [5]. In practice, the key tags are usually buried by noise tags which results in feature vectors with a flat distribution and low values.

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 leads to a shorter distance in the feature space of the PF-ART.

B. Similarity Measure

We adopt the two-way similarity measure of Fusion ART [9] for selecting the best matching category for the input image. Considering an input image img_k with its textual feature vectors $\mathbf{t}^k$, the similarity measure goes through two steps: 1) category choice and 2) winner matching. In the first step, a choice function is applied to evaluate the overall similarity between input image and each category in the category field denoted as c_j . The choice function is defined by

$$T_j = \frac{|\mathbf{t}^k \wedge \mathbf{w}_j^a|}{\alpha + |\mathbf{w}_j^a|}, \quad (1)$$

where $(p \wedge q)_i \equiv \min(p_i, q_i)$, the norm $|.|$ is defined by $|p| \equiv \sum_i p_i$, and $\alpha \approx 0$ is used to avoid the case when $|w_j| \approx 0$. The weight vector $\mathbf{w}_j^a$ is the textual feature prototype of the j^{th} cluster c_j. In the second step, the cluster having the highest value of choice function is selected as winner c_J. We use the match function to evaluate if the similarity between the feature vector of input image and prototype of c_J meets the lower bound, i.e. the vigilance parameter, which is defined by

$$M_J^a = \frac{|\mathbf{t}^k \wedge \mathbf{w}_J^a|}{|\mathbf{t}^k|} > \rho^a, \quad (2)$$

where ρ^a, handling the similarity threshold, is the vigilance parameter for the textual feature channel.

From eq.(1), we note that the similarity is the intersection between the input feature vector $\mathbf{t}^k$ and the cluster prototype $\mathbf{w}_j^a$, and the choice function assesses the degree to which the prototype $\mathbf{w}_j^a$ is a subset of the input vector $\mathbf{t}^k$. Therefore, if we interpret the feature vector using histogram, the category choice procedure selects the cluster whose interaction with the input image possesses the biggest proportion of the prototypes. However, it doesn't mean the winner category fit the input image, because if the prototypes of a given cluster are covered by the features of the input image, this category can also be chosen as winner. Therefore, the similarity measure is not symmetric. So the winner matching procedure is subsequently used to evaluate the fitness of selected cluster for the input image. The vigilance parameter ρ^a determines the lower bound for acceptance. If eq.(2) is satisfied, the input image is clustered into the winner category c_J. Otherwise, another winner category is selected from the rest of the clusters for winner matching process. If no fit category is found for the input image, a new category is generated and the prototypes are set by the visual and textual features of the input image.

We can have an further interpretation for the similarity measure. Note that textual feature indicates the presence of tags by setting the corresponding components all ones and the prototype of textual feature is modeled by the frequency distribution. Therefore the similarity measure tends to evaluate whether the input image contains key tags in the given category. The hit of more key tags means a better fit.

C. *Learning Strategy for Topic Mining*

The original learning function of Fusion ART is defined by

$$\hat{\mathbf{w}}^a_J = \beta(\mathbf{v}^k \wedge \mathbf{w}^a_J) + (1-\beta)\mathbf{w}^a_J, \tag{3}$$

where $\beta \in [0,1]$ is the learning rate and $(p \wedge q)_i \equiv \min(p_i, q_i)$. Therefore, the prototype learns from the textual feature by stably depressing the rare and unstable components while preserving the key and frequent ones. However, a set of mismatch induced by noise tags will erode the key tags of prototype. Besides, the sub-key tags cannot be preserved which may lead to the generation of extra clusters that represent the same topics. Based on the above consideration, we propose to model the cluster prototype of textual feature by the probabilistic distribution of tag occurrence. In this way, the weights of noisy tags are depressed while the key and sub-key tags can be preserved.

Consider a group of images belonging to cluster c_j, denoted as $Img^j = \{img^j_1, \ldots, img^j_l\}$. If we denote the textual feature for img^j_l as $\mathbf{t}^{j,l} = [t^{j,l}_1, \ldots, t^{j,l}_m]$ and the weight vector for textual feature of cluster c_j as $\mathbf{w}^a_j = [w^a_{j,1}, \ldots, w^a_{j,m}]$, the probability of occurrence of the k^{th} tag t^j_k of the given cluster c_j is calculated by the frequency:

$$w^a_{j,k} = p(t^j_k|c_j) = \frac{\sum_{i=1}^{l} t^{j,i}_k}{l}. \tag{4}$$

Therefore the prototype for the textual feature of cluster c_j can be represented by $\mathbf{w}^a_j = [p(t^j_1|c_j), \ldots, p(t^j_m|c_j)]$. We introduce the sequential factor and denote eq.(4) by $p_l(t^j_k|c_j)$ as the state for time l. Assuming a new image is grouped into cluster c_j, and then we can derive the relationship between time l and $l+1$ by

$$p_{l+1}(t^j_k|c_j) = \frac{\sum_{i=1}^{l+1} t^{j,i}_k}{l+1} = \frac{l}{l+1} p_l(t^j_k|c_j) + \frac{t^{j,l+1}_k}{l+1}. \tag{5}$$

Therefore, the general form of learning function for $w^a_{j,k}$ is defined by

$$\hat{w}^a_{j,k} = \frac{n_j}{n_j+1} w^a_{j,k} + \frac{t^I_k}{n_j+1}, \tag{6}$$

where n_j is the number of images in cluster c_j and t^I_k is the k^{th} component of the texture feature of input image. Consider t^I_k equals to either 0 or 1, we further simplify the learning function for $\mathbf{w}^a_j = [w^a_{j,1}, \ldots, w^a_{j,m}]$ such that

$$\hat{\mathbf{w}}^a_{j,k} = \begin{cases} \eta \mathbf{w}^a_{j,k} & t^I_k = 0 \\ \eta(\mathbf{w}^a_{j,k} + \frac{1}{n_j}) & t^I_k = 1 \end{cases}, \quad \eta = \frac{n_j}{n_j+1}. \tag{7}$$

D. *Incorporating User Preference*

Sections 4.1-4.3 describe how PF-ART works in an unsupervised manner, i.e. only the textual feature is used for clustering. In this section, the semi-supervised way that employs both textual feature as well as preference vector is presented.

With the incremental nature of Fusion ART, our method receives the user feedback by sending the user-provided relevant images as pre-defined cluster prototypes. Each component of the textual prototype (weight vector) is derived from its frequency of occurrence. The preference vector I^b, in Fig.2, is a channel for encoding user-provided labels. The preference vector does not contribute to the clustering process, but makes a tradeoff for the pre-defined categories, because of two possible cases that may decrease the clustering performance: 1) for two categories that are equal in textual vector, user may give different labels; 2) conversely, for two categories that are different in textual feature, user may give them the same label. For the first case, we combine the user-provided labels and merge them into one category. For the second case, we deem that the two categories of images are the same and represent them in one category whose textual prototype is calculated by the frequency of occurrence. Besides, the user-provided labels represent the key topics of the pre-defined categories and contribute to the generation of semantic hierarchy.

Beside the relevance feedback, users can also have a direct control of the clustering results by changing the value of vigilance parameter ρ^a in eq.(2). As illustrated, the vigilance parameter constrains the dissimilarity between the images in the same category. As the similarity of textual features directly reflects the common topics, a low vigilance parameter results in a few clusters whose key topics are few and general. In contrast, a high value leads to the generation of relatively more clusters such that the clusters belonging to one general topic are also discriminated due to detailed sub-topics. The complete algorithm of PF-ART is summarized as follows.

The clustering algorithm of PF-ART

1) Receive user preference for generating pre-defined clusters as initial network. If no user preference, create an uncommitted category with all weight vectors equal to 1.
2) Given an input image, present its textual feature vector I^a into the input field.
3) For each category c_j in category field F_2, calculate the choice function T_j (eq.(2)).
4) Extract the winner c_J such that $T_J = \max_{c_j \in F_2}\{T_j\}$.
5) Calculate the match function M^a_J (eq.(3)).
6) If $M^a_J < \rho^a$, set $T_J = 0$ and go to 3; else, go to 7.
7) If the selected c_J is uncommitted, set $\mathbf{w}^a_J = \mathbf{I}^a$ and create a new uncommitted node; else, resonance occurs, go to 8.
8) Update the weight vector $\mathbf{w}^a_J$ (eq.(7)).
9) If all images have been presented, clustering stops. Otherwise, go to 2.

V. Semantic Hierarchy Generation

After the clustering process, the key tags of each cluster are extracted as the theme lexicon representing the topics and each tag is associated with a weight indicating its importance. As the textual feature prototype represents the probability of tag occurrence, the top valued tags are extracted as key tags and weighed by their respective probability of occurrence. Besides, tags in the preference vector are all considered key tags with weights of 1s. Then, we propose an agglomerative approach for merging the clusters according to their semantic relevance.

A. *Measuring Cluster Semantic Relevance*

Given two clusters c_i and c_j, their similarity $S(c_i, c_j)$ can be expressed as the semantic relevance of key tags in their respective theme lexicons denoted as $L_i = \{l_{i,1}, \ldots, l_{i,m}\}$ and $L_j = \{l_{j,1}, \ldots, l_{j,m}\}$. Traditional measures for assessing the semantic similarities between two concepts are based on the path length according to a well-structured corpus such as WordNet [17]. But such methods are not suitable for web resources as the diversity of words used for the description of web images. Here, we follow the idea of measuring semantic similarity of two concepts based on their co-occurrence [18]. We firstly define the semantic distance between the two tags x and y. Similar to the definition of Google distance [19], the semantic distance is estimated as follows:

$$d(x,y) = \frac{\max(\log f(x), \log f(y)) - \log f(x,y)}{\log G - \min(\log f(x), \log f(y))}, \qquad (8)$$

where G is the total number of the searched images, $f(x)$ is the number of images returned by Google image search engine using keyword x and $f(x,y)$ is the number of images by searching with both x and y. Then, their semantic relevance is defined by

$$\theta(x,y) = \exp(-d(x,y)), \qquad (9)$$

where $\theta(x,y) \in [0,1]$ amd $d(x,y)$ is the semantic distance between the two tags x and y. If two concepts never occur in one image, their semantic distance becomes infinite so that their semantic relevance is 0; conversely, if two concepts always occur in one image, their semantic distance becomes 0 so that their semantic relevance is 1. Finally, we define the cluster semantic relevance between categories which can be represented as a sum of the semantic relevance between each tag in c_i and all tags in c_j weighted by the theme importance, represented by the frequency, in the respective categories:

$$S(c_i, c_j) = \sum_{r=1}^{m}\sum_{k=1}^{n} p_{i,r}p_{j,k}\theta(l_{i,r}, l_{j,k}), \qquad (10)$$

where $p_{i,r}$, equal to $w^a_{i,r}$, is the frequency of the r^{th} tag in the category c_i.

B. *Agglomerative Strategy*

In the process of having the semantic relevance $S(c_i, c_j)$ for each pair of categories, we simultaneously obtain an upper triangular matrix $v = \{v_{ij}\}$ recording the semantic relevance between pairs of categories, such that

$$v_{ij} = \begin{cases} S(c_i, c_j) & i > j \\ 0 & others \end{cases}. \qquad (11)$$

For each category c_i, we denote the set of its children categories as ξ_i. Then we define its inner scatter as:

$$\Delta_i = \max\{S(c_p, c_q) - S(c_m, c_n) | c_p, c_q, c_m, c_n \in \xi_i\}. \quad (12)$$

The merging process starts by checking if c_j is a child of c_j. Specifically, c_j is a child of c_i if and only if

$$S(c_i, c_j) + \Delta_i \geq \min S(c_p, c_q) | c_p, c_q \in \Delta_i. \qquad (13)$$

If eq.(13) is satisfied, we set c_j's father category as c_i and update the matrix using eq.(15). Otherwise, we check if c_i is a child of c_j. If both conditions are not satisfied, a new category c_{new} is generated as the father category of c_i and c_j, assigned with a new lexicon $L_{new} = \{L_i \cup L_j\}$. L_{new} contains all distinct tags in L_i and L_j. Let the k^{th} tag in L_{new} be the i^{th} tag in L_i and the j^{th} tag in L_j, its weight is determined by the following equation:

$$p_k = \frac{n_i}{n_i + n_j} p_{i,i} + \frac{n_j}{n_i + n_j} p_{j,j} = \alpha p_{i,i} + \beta p_{j,j}, \qquad (14)$$

where N_i and N_j is the number of images in c_i and c_j respectively. The equation U_i for updating the relevance score of cluster c_i in the semantic relevance matrix is defined by

$$U_i = \begin{cases} \hat{v}_{k,i} = \alpha v_{k,i} + \beta v_{k,j} & k < i \\ \hat{v}_{i,k} = \alpha v_{i,k} + \beta v_{k,j} & i < k < j \\ \hat{v}_{i,k} = \alpha v_{i,k} + \beta v_{j,k} & k > j \end{cases}, \qquad (15)$$

where α and β have the same meaning with that in eq.(14). Namely, the semantic relevance between c_{new} and other categories are the weighted average of it-s children. The algorithm is summarized as follows.

The Proposed Agglomerative algorithm

1) Construct matrix v and set the stop criterion $\bar{S}$.
2) Select the largest $S(c_i, c_j)$ in v, if $S(c_i, c_j) < \bar{S}$, algorithm stops; else go to 3.
3) Check if c_j is a child of c_i according to eq.(13), if satisfied, set c_j as a child of c_i and go to 5; else check if c_i is a child of c_j. If satisfied, set c_i is a child of c_j and go to 6; else go to 4.
4) Merge c_i and c_j into c_{new} by merging L_i and L_j into L_{new}. Set $c_i = c_{new}$. Go to 5.
5) Remove the j^{th} row and j^{th} column of v and update v by U_i (eq.(15)). Go to 2.
6) Remove the i^{th} row and i^{th} column of v and update v by U_i (eq.(15)). Go to 2.

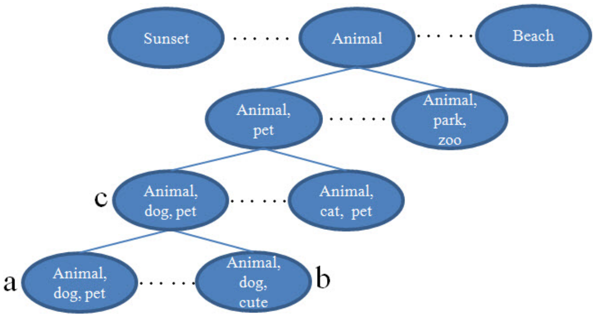

Fig. 3. A toy example for the generation procedures of semantic hierarchy.

We illustrate the merging process using a simplified case of the resulting hierarchy shown in Fig.3. The clusters a and b at the bottom are generated by PF-ART. As their key tags share "animal" and "dog", they are likely to be merged into one cluster. The cluster c is the father cluster of a set of semantically related clusters including clusters a and b. Relevant clusters are merged into a father cluster and we finally obtain a series of clusters with distinct general themes. The leaf categories should be of more constrained semantics than their father categories and the categories with the same father category should have at least one common general theme. For the visualization purpose, top tags of each cluster can be extracted as the cluster name.

VI. Experiments

We have conducted experiments on the NUS-WIDE and Flickr data sets to evaluate the performance of the proposed two-step hierarchical clustering method in three aspects: 1) the clustering quality of Probabilistic Fusion ART (PF-ART); 2) the quality of the semantic hierarchy generated by the cluster merging strategy; and 3) time cost of the whole algorithm.

A. Evaluation Measures

With the understanding that a high quality cluster maximizes the number of images of the same class in it and an ideal clustering is to group all images of the same class into one cluster, we use precision and F-score score to evaluate the quality of clusters. F-score is defined as

$$F = \frac{2(recall * precision)}{recall + precision}.$$

It has an overall assess of the quality of a cluster where a high value indicates a high quality clustering in terms of both precision and recall. Besides, as our goal is to discover the key topics of groups of images, we also evaluate the quality of clusters through their cohesion and scatter in terms of key tags, which are assessed by the cluster entropy and class entropy [20]. The cluster entropy of a cluster c_j is computed by

$$e_{c_j} = -\sum_j \frac{n_{(l_i,c_j)}}{\sum_i n_{(l_i,c_j)}} \log \frac{n_{(l_i,c_j)}}{\sum_i n_{(l_i,c_j)}}, \tag{16}$$

where $n_{(l_i,c_j)}$ denotes the number of patterns in cluster c_j with key tag l_i. It evaluates how well a cluster distinguishes images with different topics/key tags. If all of the patterns in a cluster having the same key tags, the cluster entropy is zero.

On the other hand, the class entropy evaluates whether the images having the same topics are represented by a minimal number of clusters. For each key tag l_i in cluster c_j, its class entropy is calculated by

$$\bar{e}_{l_i} = -\sum_i \frac{n_{(l_i,c_j)}}{\sum_j n_{(l_i,c_j)}} \log \frac{n_{(l_i,c_j)}}{\sum_j n_{(l_i,c_j)}}. \tag{17}$$

The overall class entropy of cluster c_j is obtained by averaging the class entropies of all the key tags. A low value of class entropy indicates a high recall of images of the same topics.

B. NUS-WIDE Data Set

The NUS-WIDE data set [21] consists of 269,648 images and 81 concepts for ground-truth. The images are downloaded from the famous photo sharing website Flickr.com. We choose this data set because it is the largest well-annotated web image set with filtered surrounding text. To test the clustering performance of our method on large scale image set, we collected 23,284 images of nine categories, including dog (2,504 images), bear (1,271 images), cat (2,376 images), bird (3,780 images), flower (3,000 images), lake (3,000 images), sky (3,000 images), sunset (3,000 images) and wedding (1,353 images), according to its ground truth labels. We choose the nine categories as they are widely used in research works and they are also the most popular tags recorded by Flickr.com.

We construct the texture feature vector by considering all distinctive and high frequency tags. Specifically, we extract all words in the raw text of selected images. After removing the stop-words, misspellings and personalized words, and stemming the variation form of words, there are 3,684 remaining tags. We further filter the infrequency tags which are not in the top 2,000 tags sorted by the tag frequency. Finally, we obtain 1,142 tags (features) and each image is associated with seven tags on average.

1) Performance of Probabilistic Fusion ART: We compare the performance of PF-ART with existing widely used text clustering algorithms including Fuzzy ART [9], K-means clustering algorithm, Isoperimetric Co-clustering Algorithm (ICA) [6], Non-negative Matrix Factorization (NMF) [7] and Semi-supervised NMF (SS-NMF) [8]. Note that the Fusion ART with one input channel is a Fuzzy ART. We compare PF-ART with Fuzzy ART to investigate the effectiveness of the learning function of PF-ART. All algorithms are implemented by C language and experiments are performed on the computer with Intel Core2 Duo CPUs 2.66GHz and 3.25GB of RAM.

To initialize PF-ART and Fuzzy ART, we fix the choice parameter α^a at 0.01 and the learning parameter β^a at 0.6. We test their performance on the nine categories of NUS-WIDE data set in terms of average precision (AP), F-score, the overall cluster entropy e and the overall class entropy entropy $\bar{e}$. For calculating the entropies, the key tags of each cluster are extracted according to top occurrence frequencies. The performance on each category is obtained by averaging the performance of its key clusters in which the number of images of the category is the majority. The overall performance is calculated by averaging the performance of all clusters of the hierarchy. As the clustering results of all the above algorithms depends on a fixed parameter, like the vigilance parameter of PF-ART and Fuzzy ART, the iteration threshold for ICA and the number of generated clusters of K-means, NMF and SS-NMF. We get the final results by averaging the performance under different settings. Specifically, the vigilance parameter ρ^a of PF-ART and Fuzzy ART is set from 0.4 to 0.9, and the iteration of ICA for bi-partitioning and the number of clusters of K-means, NMF and SS-NMF is set from 9 to 20. For the semi-supervised version of PF-ART and SS-NMF, three images of each category are used for relevance feedback. For a fair comparison, no labels are provided for PF-ART.

The results are shown in Table I and the best results are bolded. The PF-ART outperforms others in terms of average precision, F-score and cluster entropy in both unsupervised and semi-supervised cases. The clustering quality of PF-ART has a great improvement after receiving relevance feedback.

TABLE I
The performance comparison for short text clustering on NUS-WIDE data set.

NUS-WIDE	K-means	ICA	NMF	SS-NMF	Fuzzy ART	PF-ART (unsupervised)	PF-ART (semi-supervised)
AP	0.6859	0.7947	0.7412	0.8327	0.7739	0.7832	**0.8636**
$F-score$	0.5748	0.6823	0.6175	0.6917	0.6573	0.7391	**0.7624**
e	0.5882	0.4426	0.4794	0.4027	0.3842	0.3614	**0.3350**
$\bar{e}$	0.4834	0.4177	0.4136	**0.3729**	0.4364	0.3826	0.3764

TABLE II
The pformance of PF-ART and other hierarchical clustering algorithms on NUS-WIDE data set.

NUS-WIDE	HTC	HC_SL	HC_AL	HC_CL	HFC	HCC	PHTC
AP	0.6692	0.5977	0.6512	0.5485	0.6309	0.7248	**0.7634**
$F-score$	0.4977	0.5291	0.5047	0.4628	0.4811	0.4631	**0.5883**
e	0.4642	0.4726	0.4873	0.5581	0.5385	0.4468	**0.4434**
$\bar{e}$	0.5258	0.5873	0.5131	0.6127	0.4871	0.5235	**0.4604**
e_{max}	0.5471	0.6272	0.6894	0.7284	0.5813	0.5707	**0.5137**
$\bar{e}_{max}$	0.6963	0.7642	0.6826	0.7535	0.6427	0. 7364	**0.6355**
$Time(sec.)$	108.1504	165.5269	182.8592	151.1495	136.4930	86.1498	**32.2217**

TABLE III
The performance comparison for short text clustering on Flickr data set.

Flickr	K-means	ICA	NMF	SS-NMF	Fuzzy ART	PF-ART (unsupervised)	PF-ART (semi-supervised)
AP	0.7559	0.8347	0.8147	0.8793	0.8439	0.8363	**0.8812**
$F-score$	0.6644	0.7025	0.6892	0.7636	0.7350	0.7731	**0.7848**
e	0.3022	0.2491	0.2685	0.2106	0.2411	0.2317	**0.1934**
$\bar{e}$	0.4454	0.4136	0.4329	0.3801	0.4284	0.4013	**0.3704**

TABLE IV
The performance of PF-ART and other hierarchical clustering algorithms on Flickr data set.

Flickr	HTC	HC_SL	HC_AL	HC_CL	HFC	HCC	PHTC
AP	0.7269	0.6168	0.6823	0.5578	0.6581	0.7155	**0.8128**
$F-score$	0.5366	0.5594	0.4989	0.5168	0.5267	0.4814	**0.6933**
e	0.3942	0.4462	0.4131	0.5083	0.4325	0.3768	**0.2434**
$\bar{e}$	0.4712	0.4296	0.4328	0.4056	0.4203	0.4648	**0.3869**
e_{max}	0.5714	0.6420	0.6359	0.6821	0.5329	0.4966	**0.4137**
$\bar{e}_{max}$	0.6424	0.6342	0.6593	0.6684	0.6341	0. 6139	**0.5944**
$Time(sec.)$	41.8445	48.4286	59.8663	51.3578	46.8524	36.3776	**22.9314**

Compared with Fuzzy ART, PF-ART without relevance feedback gets similar result of precision and cluster entropy, but has a better score in F-score and class entropy, which indicates a higher recall. The reason should be that Fuzzy ART cannot preserve sub-topics. As with a high vigilance parameter, more clusters are generated due to the mismatch of sub-topics. SS-NMF obtains the best result in class entropy. One possible reason for PF-ART is still the side-effect of noisy tags which increases the difference between the cluster prototype and the input pattern. However, the performance of PF-ART is still comparable to the best result.

2) Performance of the proposed PHTC: As the proposed two-step Personalized Hierarchical Theme-based Clustering approach (PHTC) is an agglomerative clustering algorithm in nature. We compare its performance with four related methods. The first method, referred to as hierarchical theme-based clustering (HTC), directly applies our merging strategy on the input patterns without the clustering step. Specifically, each image is regarded as one cluster with the associated tags as key tags, and then the cluster semantic evaluation and merging strategies are performed to obtain the semantic hierarchy. The second method is the traditional agglomerative method (HC). As different merging strategies vary largely on performance, we test three popular merging strategies termed single-linkage (HC-SL) [12], average-linkage (HC-AL) [13] and complete-linkage (HC-CL) [14]. The third and last methods are Hierarchical Fuzzy Clustering (HFC) [22] and the hierarchical clustering algorithm (HCC) used in Hierarchical Comments-based Clustering [4].

The parameter setting of PF-ART in PHTC follows that in section 6.2.1. The vigilance parameter ρ^a is fixed at 0.9 and no relevance feedback is provided to PF-ART. HFC requires two parameters including the number of nearest neighbors k and the number of clusters s for stop criterion. We empirically set $k = 100$ and $s = 19$ according to the size and the number of topics of our data set. To make a fair comparison, the stop criterion $s = 19$ is applied to all the algorithms.

We evaluate the quality of the generated hierarchy by evaluating the quality of all clusters in the hierarchy. Besides the overall cluster and class entropies e and $\bar{e}$, the maximum entropies e_{max} and $\bar{e}_{max}$ are given for revealing the worst merging in the hierarchy. We also consider the time cost for generating the hierarchy. The results are shown in Table II, from which we observe that PHTC obtains the best results for all evaluation measures. Compared with HTC, the quality of generated hierarchy has a great improvement in terms of average precision and F-score. Note that HTC is a special case of PHTC when $\rho^a = 1$, it demonstrates the effectiveness of PF-ART in grouping semantically similar images and mining the key tags. We also observe that the performance of HTC is comparable to the best results of other algorithms. It indicates that the proposed cluster semantic relevance measure and merging strategy are effective for text clustering. In terms of the time cost, PHTC is much faster than other algorithms. It is due to the rapid nature of PF-ART. Thus, our algorithm is suitable for large image data sets.

C. Flickr Data Set

To evaluate the robustness of our algorithm for universal web image resources, we conduct experiments on another image set also crawled from Flickr.com. Although the two data sets are collected from the same website, the images are totally different as there is a long interval between the collections of these two data sets. This data set contains 11,589 images of 19 categories (animal, baby, beach, birthday, boat, crowd, graduation, museum, night, parade, park, people, picnic, playground, show, skiing, sport, sunset and wedding) and each image is associated with filtered textual description (i.e. tags). Therefore, they can be seen as two different resources. In total, there are 894 tags (features) and each image is associated with six tags on average.

Similar to the experiments on NUS-WIDE data set, PF-ART with relevance feedback, in Table III, achieves the best performance in terms of all evaluation criteria and has an great improvement, compared with the unsupervised one, on average precision. We can also observe that PF-ART without relevance feedback obtain comparable performance with the best results of other unsupervised methods. Table IV shows similar results with that observed in Table II. PHTC outperforms other hierarchical methods in both clustering quality and time cost. Interestingly, it is shown that the number of images of the NUS-WIDE is twice as much as that of the Flickr data set while the time cost is only one and a half times. It demonstrates that the mining of semantic groups can enhance the quality of generated hierarchy and reduces the computational cost. Therefore, PHTC is scalable and efficient for the large image collections. A snapshot of the resulting hierarchy in the experiment is shown in Fig.4. Each folder denotes a cluster and the folder name includes the top tags of that cluster. A better interface can be achieved by simple post-processing such as the name pruning of sub-clusters.

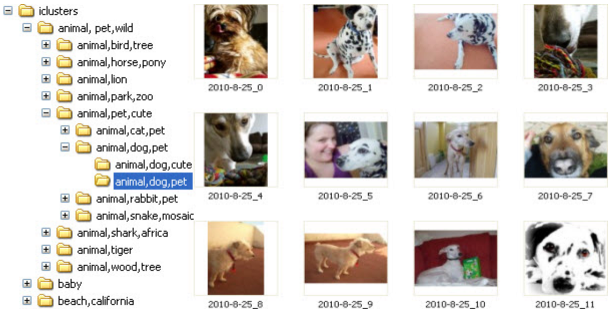

Fig. 4. A snapshot of the generated hierarchy on Flickr data set.

VII. CONCLUSION

In this paper, we propose a two-step semi-supervised hierarchical clustering algorithm for web image organization. The main contributions of our proposed framework include: 1) a novel clustering algorithm Probabilistic Fusion ART (PF-ART) which is effective and rapid for short text clustering and semi-supervised learning; 2) an effective cluster semantic relevance measure (CSR) and merging strategy for associating semantically similar clusters. Different from typical agglomerative algorithms, our approach can organize the clusters into a multi-branch tree structure, which provides a more compact and systematic interface for image organization. There still remains several issues to be solved: 1) the filtering methods for noise tags; and 2) the parameter dependence of the clustering model. Therefore, our future work will focus on the effective methods for extracting tags from original surrounding text and the improvement of the proposed PF-ART clustering algorithm.

ACKNOWLEDGMENT

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

REFERENCES

[1] F. Jing, C. Wang, Y. Yao, L.Zhang, and W. Ma, "Igroup: web image search results clustering," *In Proc. ACM Multimedia*, pp. 377–384, 2006.

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

[3] H. Ding, J. Liu, and H. Lu, "Hierarchical clustering-based navigation of image search results," *In Proc. ACM Multimedia*, pp. 741–744, 2008.

[4] C. Hsu, J. Caverlee, and E. Khabiri, "Hierarchical comments-based clustering," *In Proc. ACM SAC*, pp. 1130–1137, 2011.

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

[6] M. Rege, M. Dong, and F. Fotouhi, "Co-clustering documents and words using bipartite isoperimetric graph partitioning," *In Proc. ICDM*, pp. 532–541, 2006.

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

[8] Y. Chen, M. Rege, M. Dong, and J. Hua, "Incorporating user provided constraints into document clustering," *In ICDM*, pp. 103–112, 2007.

[9] 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.

[10] Y. Chen, M. Dong, and W. Wan, "Image co-clustering with multi-modality features and user feedbacks," *In MM*, pp. 689–692, 2007.

[11] X. Shi, W. Fan, and P.S. Yu, "Efficient semi-supervised spectral co-clustering with constraints," *In Proc. ICDM*, pp. 532–541, 2010.

[12] J.C. Gower and G.J.S. Ross, "Minimum spanning trees and single linkage clustering analysis," *J. R. Stat. Soc. Ser. C*, pp. 595–616, 1969.

[13] H. Schtze and C. Silverstein, "Projections for efficient document clustering," *In proc. SIGIR*, pp. 74–81, 1997.

[14] O. Aichholzer and F. Aurenhammer, "Classifying hyperplanes in hypercubes," *SIAM J. Discrete Math.*, pp. 225–232, 1996.

[15] 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.

[16] A.-H. Tan, H.-L. Ong, H. Pan, J. Ng, and Q Li, "Towards personalized web intelligence," *Knowl. Inf. Syst.*, pp. 595–616, 2004.

[17] T. Pedersen, S. Patwardhan, and J. Michelizzi, "Wordnet::similarity: measuring the relatedness of concepts," *Demonstration papers at HLT-NAACL*, 2004.

[18] D. Liu, X. Hua, L. Yang, M. Wang, and H. Zhang, "Tag ranking," *In Proc. WWW*, pp. 351–360, 2009.

[19] R. Cilibrasi and P. M. B. Vitanyi, "The google similarity distance," *In TKDE*, pp. 370–383, 2007.

[20] J. He, A.-H. Tan, C.-L. Tan, and S-Y. Sung, "On quantitative evaluation of clustering systems," *Clustering and Information Retrieval, Kluwer Academic Publishers*, pp. 105–133, 2003.

[21] 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 proc. CIVR*, pp. 1–9, 2009.

[22] L. Li and Y. Liang, "A hierarchical fuzzy clustering algorithm," *In proc. ICCASM*, pp. 248–255, 2010.