2011 International Conference on Document Analysis and Recognition

Web Multimedia Object Clustering via Information Fusion

Wenting Lu[1,2], Lei Li[2], Tao Li[2]
[1]*Pattern Recognition and Intelligence System Lab*
Beijing University of Posts and Telecommunications
Beijing 100876, P.R.China
Email: {wlu,lli003,taoli}@cs.fiu.edu

Honggang Zhang[1], Jun Guo[1]
[2]*School of Computing and Information Sciences*
Florida International University
Miami, Florida 33199, United States
Email: {zhhg,guojun}@bupt.edu.cn

***Abstract*—Multimedia information plays an increasingly important role in humans daily activities. Given a set of web multimedia objects (images with corresponding texts), a challenging problem is how to group these images into several clusters using the available information. Previous researches focus on either adopting individual information, or simply combining image and text information together for clustering. In this paper, we propose a novel approach (*Dynamic Weighted Clustering*) to separate images under the "supervision" of text descriptions; Also, we provide a comparative experimental investigation on utilizing text and image information to tackle web image clustering. Empirical experiments on a manually collected web multimedia object (related to the events after disasters) dataset are conducted to demonstrate the efficacy of our proposed method.**

***Keywords*-Multimedia Object Clustering, Image, Text, Information Fusion, Dynamic Weighting**

I. Introduction

Multimedia information plays an increasingly important role in humans daily activities. With the rapid development of technologies of fast access to the Internet and the popularization of digital cameras, enormous digital images are posted and shared online everyday. Besides great convenience, how to efficiently search images that satisfy the need of web users in large-scale multimedia databases is becoming more and more challenging. To tackle this problem, an intuitive way is to organize the image databases by grouping the images into different clusters, and then performing the subsequent processes, *e.g.*, image indexing, on the generated image clusters. Therefore, web image clustering, as one of the most crucial steps in web image retrieval, attracts increasing attention in recent years.

Web image clustering is a non-trivial task in terms of the diversity of image content and the limited available information. Fortunately, text information is often provided by web users to describe the general contents of images, *e.g.*, image titles, headers, or text descriptions assigned to them. An alternative solution to web image clustering is to integrate such text information with image information to generate a hybrid information space. Specifically, we can initially extract different types of features (image features and text features), and then group the images by analyzing the special characteristics of the integrated version of different features. By doing this, the distinguishable power of low-level visual information extracted from images can be enhanced.

In this paper, we explore the feasibility of using text information as a "guidance" for image clustering by proposing a novel method – *Dynamic Weighted Clustering*. Specifically, the proposed approach assumes that different image features might carry different significance for representing images, and under the "supervision" of text features, we can **dynamically decide the importance of different image features**. It is straightforward that the important features may have more distinguishable power for the similarity computation. By doing this, a weighted similarity measurement can be formalized and applied to calculating the pairwise similarities for images. Further, we employ *Symmetric Nonnegative Matrix Factorization* on the similarity matrix, and finally achieve better clustering performance. Moreover, we provide a comparative experimental study on the issue of integrating text and image information together for clustering. Empirical experiments on a manually collected web multimedia object (related to the events after disasters) dataset are conducted to demonstrate the efficacy of our proposed method.

The rest of this paper is organized as follows. In Section II we review some related works of information fusion in multimedia object clustering. The algorithmic detail of our proposed approach is given in Section III. Section IV presents a comprehensive experimental comparison among different clustering approaches and finally we conclude the paper in Section V.

II. Related Work

Most of the traditional web image clustering approaches [1], [2], [3] often focus on utilizing individual information to group images. Due to the simplicity of feature extraction and computation, these techniques are still adopted by the majority of web image applications. However, some problems along with image representations arise in different scenarios. For text-based image clustering, (1) it is difficult to group images if there is no text information assigned to them; (2) the manual text labeling is too subjective due to human assignments. For image feature based clustering, (1) the extracted image features tend to

1520-5363/11 $26.00 © 2011 IEEE
DOI 10.1109/ICDAR.2011.72

be substantial, making clustering suffering in high dimensional space; (2) low-level visual features cannot represent semantic meanings, and hence the distinguishable power of image features is relatively poor when performing clustering task. Therefore, the performance of traditional web image clustering approaches is very limited.

To solve the above problems, many research publications [4], [5] design multi-source clustering algorithms to group images. In general, multi-source clustering methods can be categorized into three different groups:

- Feature Integration: Enlarge the feature representation to incorporate all attributes from different sources and produce a unified feature space. The advantage of feature integration is that the unified feature representation is often more informative and also allows many different data mining methods to be applied and systematically compared. One disadvantage is the increased computational complexity and difficulty as the data dimension becomes large [6];
- Similarity Integration: Keep the feature spaces in their original form and integrate them at the similarity computation or the kernel level. Different weights can be used for different data sources. Standard clustering methods can then be applied once the total similarity is computed.
- Consensus Clustering: Keep data intact in their original form and apply clustering methods to each feature space separately. Clustering results on different feature spaces are then combined via consensus clustering algorithms [7].

Our contribution: In this paper, we first investigate the feasibility of different clustering algorithms on text and image representations mentioned above. Specifically, we compare some well-known clustering approaches, such as partition-based algorithms (*K-means*), agglomerative algorithms (*Hierarchical Clustering* with single-link, complete-link, average-link and ward), matrix-based methods (*Non-negative Matrix Factorization* [8]) and graph-based algorithms (*Spectral Clustering* [9]). In addition, we empirically explore the efficacy of combining different clustering results (namely *Consensus Clustering* [10]) to generate a consensus clustering result. Furthermore, we propose a novel clustering method – *Dynamic Weighted Clustering*, in which text information is regarded as a "guidance" for image clustering. Experimental results demonstrate the effectiveness of our proposed method.

III. Dynamic Weighted Clustering

Web image clustering is prerequisite for web-based multimedia applications and has a direct impact to the speed and accuracy of these applications. However, all the approaches mentioned in Section II focus on simply combining two data sources (text and image information), and none of them takes the advantage that one data source can provide "guidance" for another on how to perform clustering task.

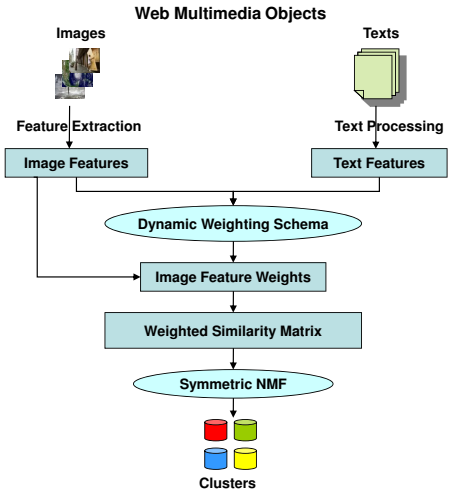

Figure 1. The framework of *Dynamic Weighted Clustering*.

In this paper, besides the empirical comparative study of the previous works, we also propose a novel web image clustering method – *Dynamic Weighted Clustering* – which employs the text-based information (*i.e.*, the texts relevant to the images) to guide the procedure of dividing images into different groups, and consequently achieve better web multimedia object clustering results. Fig 1 presents the framework of our proposed method.

A. Dynamic Weighting Schema

Image feature extraction techniques tend to extract a huge number of image features based on different criteria. Among these features, some of them might carry significant semantic information about the image, whereas some others might be less crucial. Particularly in web image clustering, the extracted features should be more representative and carry more significance when calculating the object pairwise similarities. Therefore, it might be helpful for image clustering to learn a better similarity measurement by dynamically assigning different weights to different image features so that the features with more importance can be captured and play more meaningful roles on grouping images. Our previous works [11] on music information retrieval showed how to learn appropriate similarity metrics based on the correlation between acoustic features and user access patterns of music, and [12] showed the efficacy of Dynamic Weighting in supervised learning. Motivated by these, we integrate the concept of dynamic feature weighting into our solution to clustering which is the unsupervised learning.

Specifically in web image clustering, given that human perception of an image is well approximated by its text description, a good weighting schema for the extracted image features guided by text information may lead to a high-quality similarity measurement, and therefore better

clustering results. Let $\mathbf{m}_i = (\mathbf{f}_i, \mathbf{t}_i)$ denote the i-th image in the image collection, where $\mathbf{f}_i$ and $\mathbf{t}_i$ represent its image features and text features respectively. Let $S_f(\mathbf{f}_i, \mathbf{f}_j; \mathbf{w}) = \sum_l f_{i,l} f_{j,l} w_l$ be the image-based similarity measurement between the i-th and the j-th images when the parameterized weights are given by $\mathbf{w}$. Let $S_t(\mathbf{t}_i, \mathbf{t}_j) = \sum_k t_{i,k} t_{j,k}$ be the similarity measurement between the i-th and the j-th images based on their text description features, in general, the words with specific meanings extracted from text. Here for each k, $t_{i,k}$ denotes whether the k-th word appears in the text description of the i-th image. To learn appropriate weights $\mathbf{w}$ for image features, we can enforce the consistency between similarity measurements $S_f(\mathbf{f}_i, \mathbf{f}_j; \mathbf{w})$ and $S_t(\mathbf{t}_i, \mathbf{t}_j)$. The above idea leads to the following optimization problem:

$$\mathbf{w}^* = argmin \sum_{i \neq j} (S_f(\mathbf{f}_i, \mathbf{f}_j; \mathbf{w}) - S_t(\mathbf{t}_i, \mathbf{t}_j))^2 \quad s.t. \mathbf{w} \geq 0. \tag{1}$$

By rewriting and calculating the summation in Eq.(1), the optimization problem can be addressed using quadratic programming techniques [13]. For detailed mathematical deduction, please refer to [11]. After obtaining the optimal weighting information for each image feature, we can get a weighted similarity metric for pairwise similarity calculation when performing clustering task. Specifically, assume that we have two image objects $\mathbf{f}_i$ and $\mathbf{f}_j$, represented by their image-based features respectively, and the optimal weighting schema $\mathbf{w}^*$, the similarity between these two images can be computed as

$$Sim(\mathbf{f}_i, \mathbf{f}_j) = 1 - \sqrt{\sum_l w_l^* \times (f_{i,l} - f_{j,l})^2} \tag{2}$$

where $\mathbf{f}_i$, $\mathbf{f}_j$ and $\mathbf{w}^*$ are normalized, and $\mathbf{w}^*$ s.t. $\sum_l w_l^* = 1$, so that each similarity score is in the range of [0,1]. Here we use Euclidean distance as the base similarity measurement.

Based on this weighted similarity measurement, we can obtain a pairwise similarity matrix M on the entire web multimedia object collection, where each entry M_{ij} represents the pairwise similarity between the i-th image and the j-th image. It is easy to observe that the matrix M is a symmetric matrix, and the elements on its diagonal all equal to 1. In the following, we will employ *Symmetric Nonnegative Matrix Factorization* [14] (*SNMF* for short) on this similarity matrix to solve the clustering problem.

B. Symmetric NMF on Weighted Similarity Matrix

Once we obtain the similarity matrix, clustering algorithms need to be performed to group these images into clusters. Most matrix-based clustering algorithms deal with a rectangular data matrix (*e.g.*, document-term matrix and sentence-term matrix in text mining), and they are not suitable for clustering a pairwise similarity matrix. In our work, we adopt the *SNMF* algorithm to conduct the clustering.

Given a pairwise similarity matrix M, we want to find H such that

$$\min_{H \geq 0} J = ||M - HH^T||^2, \tag{3}$$

where the matrix norm $||X||^2 = \sum_{ij} X_{ij}^2$ is the Frobenius norm. To derive the updating rule for Eq.(3) with nonnegative constraints, $H_{ij} \geq 0$, we introduce the Lagrangian multipliers λ_{ij} and let $L = J + \sum_{ij} \lambda_{ij} H_{ij}$. The first order KKT condition for local minima is

$$\frac{\partial L}{\partial H_{ij}} = \frac{\partial J}{\partial H_{ij}} + \lambda_{ij} = 0, \text{ and } \lambda_{ij} H_{ij} = 0, \forall i, j.$$

Note that $\frac{\partial J}{\partial H_{ij}} = -4MH + 4HH^TH$. Hence the KKT condition leads to the fixed point relation:

$$(-4MH + 4HH^TH)_{ij} H_{ij} = 0. \tag{4}$$

Using gradient descent method, we have

$$H_{ij} \leftarrow H_{ij} - \epsilon_{ij} \frac{\partial J}{\partial H_{ij}}. \tag{5}$$

By setting $\epsilon_{ij} = \frac{H_{ij}}{(8HH^TH)_{ij}}$, we can obtain the NMF style multiplicative updating rule for *SNMF*:

$$H_{ij} \leftarrow \frac{1}{2}[H_{ij}(1 + \frac{(MH)_{ij}}{(HH^TH)_{ij}})]. \tag{6}$$

Theorem 1: **The loss $||M - HH^T||^2$ is non-increasing under the update rule given by Eq.(6).**

One can use the strategy similar to [8] for the proof of Theorem 1. Based on the above analysis, the algorithm procedure for solving *SNMF* is: give an initial guess of H, iteratively update H using Eq.(6) until convergence. This gradient descent method will converge to a local minima of the problem.

It has been shown that *SNMF* is equivalent to kernel K-means clustering and is a special case of tri-factor NMF [15]. Another important property is that the simple *SNMF* is equivalent to the sophisticated normalized cut spectral clustering. These results demonstrate the clustering ability of *SNMF*. Specifically in our proposed clustering method, *SNMF* is adopted to perform clustering on the weighted pairwise similarity matrix. The resulted matrix H contains the cluster indicator for each object. In the experiments, we will show how much the clustering results be influenced by our dynamic weighting schema.

IV. Experimental Results

A. Real World DataSet

In order to empirically study the previous multimodal feature combination methods and compare them with our proposed method, we collected 355 web multimedia objects (images with corresponding texts) about "the aftermath of disasters", which include 4 different topics: Hurricane_building_collapse, Hurricane_flood, Oil_spill_seagrass, and Oil_spill_Animal_death. Each topic

includes 101, 101, 53 and 100 objects respectively. Unlike the traditional scene object databases, which are mainly focusing on visual categorization, web images are usually organized by topics of events and each image might reflect only one aspect of the whole event, such that images in the same group may vary visually but very similar in terms of semantic concepts.

B. Design of Experiments

In our experiment, web objects with different feature spaces are regarded as the input to various clustering algorithms. For text feature extraction, we initially analyze each text description related to each image, and then obtain some original high-frequency terms in these texts by using MALLET [16]. In order to represent images effectively, we adopt CEDD (Color and Edge Directivity Descriptor) [17], which is a new feature descriptor incorporating both color feature and texture feature. In CEDD, a novel and effective method is adopted to integrate a 24-bins color histogram and a 6-bins texture histogram to form a final 144-bins histogram. One of the most important characteristics of CEDD is its low computational power needed for feature extraction, in comparison with the needs of most of MPEG-7 descriptors. The feature spaces are described as follows:

- Text-based (*Text* for short): Note that the image feature is a 144-dimension vector, while the cardinality of the text features is 1788. To balance the contribution of different features to the clustering results, we choose the top 144 terms with high frequency as the text features;
- Image-based (*Img* for short): Extract 144-dimension CEDD features from the images to perform clustering;
- Feature Integration (*Feat* for short): Combine the above two features together by simply concatenating them to form a 288-dimension vector;
- Feature Selection on Feat (*PCA* for short): Based on the feature integration, perform *Principal Component Analysis* to reduce the dimension of the integrated feature vector from 288 to 144;
- Similarity Integration (*Sim* for short): Compute the pairwise similarity using text-based features and image-based features (CEDD) respectively, and then use the weighted summation of these two types of similarities as the similarity measurement between objects. Based on experiments, the optimal weight factor is 0.5.

After obtaining different feature space representations for objects, we compare multiple clustering algorithms mentioned in Section II to investigate the effect of different representations. Since the results of some algorithms are nondeterministic, we run these algorithms 10 trials and take the average accuracy as the final results. Here the accuracy is calculated based on the *true positive* and *true negative* of image pairs.

C. Experiment Results

Table I shows the clustering accuracy comparisons of different clustering algorithms on the investigated feature space representations. From the comparison results, we have the following empirical observations:

Table I
EXPERIMENTAL COMPARISON OF CLUSTERING.

	K-means	Hierarchical				NMF	Spectral
		Single	Complete	Average	Ward		
Text	0.5211	0.3465	0.5972	**0.6310**	0.5944	0.5963	0.5865
Img	0.3786	0.2901	0.3662	0.3268	0.3268	0.3949	**0.5236**
Feat	0.6268	0.3211	0.6141	0.5746	**0.6648**	0.6321	0.6250
PCA	0.3377	0.3239	0.3211	0.2901	0.5831	0.5921	**0.6310**
Sim	0.5321	0.3268	0.6338	0.5887	**0.6648**	0.6539	0.6253

1) The single usage of text features achieves better performance than image features in most clustering algorithms.
2) By simply combining two different data sources as the fundamental base, such as *Feat* and *Sim*, the accuracy of clustering results can be improved in most cases. Such multi-modal fusion techniques demonstrate that the text information can provide "guidance" for clustering image collection to some extent.
3) The average performance of *PCA* is relatively poor, compared with other representations. The intuitive explanation is that *PCA* regards each feature in the integrated feature space as from the same source; however, different features from multiple sources in the integrated version might carry distinct importance when calculating the pairwise similarities of objects.

For further comparison, we explore the possibility of integrating different clustering results via consensus clustering [10]. Specifically, we first formulate the consensus clustering as a nonnegative matrix factorization problem, and then randomly select 5 different class assignments each time from the clustering results mentioned above. Consensus clustering is performed on the selected cluster assignments for 10 times. For our proposed method, *Dynamic Weighted Clustering* (*DyW* for short), we also perform the clustering 10 trials, and take the average accuracy as the experimental result. Finally we compare the average accuracy of our proposed method with the best experimental result of different feature space representations, along with the average result of the consensus clustering.

The comparison is presented in Fig 2. From the result, we observe that the average performance on accuracy of our proposed method is better than the best results of the previous methods and is comparable to the performance of consensus clustering (*Ensm* for short). Here, consensus clustering achieve relatively outstanding result because it can improve clustering robustness, deal with distributed and heterogeneous data sources and make use of multiple clustering criteria. The reason for the performance improvement of our proposed method is straightforward:

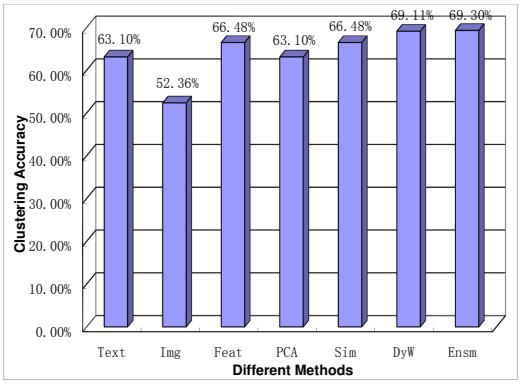

Figure 2. Comparison among the best performance of the previous methods and our proposed method.

1) *Dym* aims to employ advantages of one data source to help clustering on the other data source. In other words, the proposed method utilizes the text-based information to find out the best weighting schema for the original image features.
2) *SNMF* has several nice properties that makes it a powerful tool for clustering. One of the nice properties of the *SNMF* algorithm is its inherent ability for maintaining the near-orthogonality of H, which is important for object clustering. Since an exact orthogonality implies that each row of H can have only one non-zero element, which leads to the hard clustering of objects; while a non-orthogonality of H does not have a clustering interpretation. Therefore, the near-orthogonality conditions of *SNMF* allow for "soft clustering", *i.e*, each object can belong fractionally to multiple clusters, which usually leads to clustering performance improvement.

V. Concluding Remarks

In this paper, we empirically study the problem of combining two data sources (text and image) to perform web object clustering and show that combining two data sources can lead to better results comparing with using single data source. Moreover, our proposed *Dynamic Weighted Clustering* which can effectively utilize the image-related text information to find out better schema to group the web images into clusters. The empirical results show that our proposed method outperforms the previous methods in terms of the accuracy of clustering. For future work, we are interested in utilizing semantic information hidden in texts to facilitate clustering web objects. Also, to handle large scale web object clustering, it is better to transform our proposed method onto distributed computing framework, e.g., MapReduce [18].

VI. Acknowledgement

This work is partially supported by the U.S. Department of Homeland Security's VACCINE Center under Award Number 2009-ST-061-CI0001 and by the Army Research Office under grant number W911NF-10-1-0366, and by National Natural Science Foundation of China under Grant No.61005004, the Fundamental Research Funds for the Central Universities under Grant No.2009RC0105 and Scientific Research Foundation for the Returned Overseas Chinese Scholars, State Education Ministry. This project is also funded by Qualcomm, Inc.

References

[1] Y. Chen, J. Wang, and R. Krovetz, "Content-based image retrieval by clustering," in *Proceedings of ACM MM*. ACM, 2003, pp. 193–200.

[2] W. Sunayama, A. Nagata, and M. Yachida, "Image clustering system on WWW using Web texts," in *Proceedings of HIS*. IEEE, 2005, pp. 230–235.

[3] R. Agrawal, W. Grosky, and F. Fotouhi, "Image clustering using multimodal keywords," *Semantic Multimedia*, pp. 113–123, 2006.

[4] 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 *Proceedings of ACM MM*. ACM, 2004, pp. 952–959.

[5] 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," in *Proceedings of ACM MM*. ACM, 2005, pp. 112–121.

[6] L. Wu, S. Oviatt, and P. Cohen, "Multimodal integration-a statistical view," *IEEE Transactions on Multimedia*, vol. 1, no. 4, pp. 334–341, 2002.

[7] S. Monti, P. Tamayo, J. Mesirov, and T. Golub, "Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data," *Machine learning*, vol. 52, no. 1, pp. 91–118, 2003.

[8] D. Lee and H. Seung, "Algorithms for non-negative matrix factorization," *Advances in neural information processing systems*, vol. 13, 2001.

[9] A. Ng, M. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in *Proceeding of ANIPS*, 2001, pp. 849–856.

[10] T. Li, C. Ding, and M. Jordan, "Solving consensus and semi-supervised clustering problems using nonnegative matrix factorization," in *Prceedings of ICDM*, 2007, pp. 577–582.

[11] B. Shao, M. Ogihara, D. Wang, and T. Li, "Music recommendation based on acoustic features and user access patterns," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 8, pp. 1602–1611, 2009.

[12] L. Li, W. Lu, J. Li, T. Li, H. Zhang, and J. Gun, "Exploring Interaction Between Images and Texts for Web Image Categorization," in *Proceedings of FLAIRS*, 2011, pp. 45–50.

[13] P. Gill, W. Murray, and M. Wright, "Practical optimization," 1981.

[14] D. Wang, T. Li, S. Zhu, and C. Ding, "Multi-document summarization via sentence-level semantic analysis and symmetric matrix factorization," in *Proceedings of SIGIR*. ACM, 2008, pp. 307–314.

[15] C. Ding, T. Li, W. Peng, and H. Park, "Orthogonal nonnegative matrix t-factorizations for clustering," in *Proceedings of SIGKDD*. ACM, 2006, pp. 126–135.

[16] A. McCallum, "MALLET: A Machine Learning for Language Toolkit," 2002, http://mallet.cs.umass.edu.

[17] S. Chatzichristofis and Y. Boutalis, "Cedd: Color and edge directivity descriptor: A compact descriptor for image indexing and retrieval," in *Proceedings of ICCVS*, 2008, pp. 312–322.

[18] J. Dean and S. Ghemawat, "MapReduce: Simplified data processing on large clusters," *Communications of the ACM*, vol. 51, no. 1, pp. 107–113, 2008.