Volume 3, Issue 11, November 2013 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Affinity Propagation Clustering with Background Knowledge Using Pair Wise Constraints

K.M. Padmapriya
Asst. Professor, Dept. of Computer Science,
SSM College of Arts and Science
Komarapalayam, India

M. Sumathi.
M.Sc.,Research Scholar
SSM College of Arts and Science,
Komarapalayam, India

Abstract - The pair wise constraints specifying a pair of samples should be grouped together or not have been successfully incorporated into the conventional clustering methods such as k-means and spectral clustering for the performance enhancement. Nevertheless, the issue of pair wise constraints has not been well studied in the recently proposed maximum margin clustering (MMC), which extends the maximum margin framework in supervised learning for clustering and often shows a promising performance. This paper therefore proposes a pair wise constrained MMC algorithm. Clustering which tries to group a set of points into clusters such that points in the same cluster are more similar to each other than points in different clusters, under a particular similarity metric. In the generative clustering model, a parametric form of data generation is assumed, and the goal in the maximum likelihood formulation is to find the parameters that maximize the probability (likelihood) of generation of the data given the model. In the most general formulation, the number of clusters k is also considered to be an unknown parameter. Such a clustering formulation is called a "model selection" framework, since it has to choose the best value of k under which the clustering model fits the data. In clustering process, semi-supervised learning is a class of machine learning techniques that make use of both labeled and unlabeled data for training - typically a small amount of labeled data with a large amount of unlabeled data. Semi-supervised learning falls between unsupervised learning (without any labeled training data) and supervised learning (with completely labeled training data).

Keywords:

I. INTRODUCTION

Document clustering (also referred to as Text clustering) is closely related to the concept of data clustering. Document clustering is a more specific technique for unsupervised document organization, automatic topic extraction and fast information retrieval or filtering. A web search engine often returns thousands of pages in response to a broad query, making it difficult for users to browse or to identify relevant information. Clustering methods can be used to automatically group the retrieved documents into a list of meaningful categories, as is achieved by Enterprise Search engines

II. CLUSTERING

Clustering is an unsupervised learning problem, which tries to group a set of points into clusters such that points in the same cluster are more similar to each other than points in different clusters, under a particular similarity metric (Jain & Dubes, 1988). Here, the learning algorithm just observes a set of points without observing any corresponding class/category labels. Clustering problems can also be categorized as generative or discriminative. In the generative clustering model, a parametric form of data generation is assumed, and the goal in the maximum likelihood formulation is to find the parameters that maximize the probability (likelihood) of generation of the data given the model. In the most general formulation, the number of clusters k is also considered to be an unknown parameter. Such a clustering formulation is called a "model selection" framework, since it has to choose the best value of k under which the clustering model fits the data. We will be assuming that k is known in the clustering frameworks that we will be considering, unless explicitly mentioned otherwise. In the discriminative clustering setting (e.g., graph-1 theoretic clustering), the clustering algorithm tries to cluster the data so as to maximize within-cluster similarity and minimize between-cluster similarity based on a particular similarity metric, where it is not necessary to consider an underlying parametric data generation model. In both the generative and discriminative models, clustering algorithms are generally posed as optimization problems and solved by iterative methods like EM (Dempster, Laird, & Rubin, 1977), approximation algorithms like KMedian (Jain & Vazirani, 2001), or heuristic methods like Metis (Karypis & Kumar, 1998).

A. SUPERVISED LEARNING

Supervised learning is the machine learning task of inferring a function from supervised training data. The training data consist of a set of training examples. In supervised learning, each example is a pair consisting of an input object (typically a vector) and a desired output value (also called the supervisory signal). A supervised learning algorithm analyzes the training data and produces an inferred function, which is called a classifier (if the output is discrete, see classification) or a regression function (if the output is continuous, see regression). The inferred function should predict the correct output value for any valid input object. This requires the learning algorithm to generalize from the training data to unseen situations in a "reasonable" way (see inductive bias). (Compare with unsupervised learning.) The parallel task in human and animal psychology is often referred to as concept learning.

B. SEMI-SUPERVISED CLUSTERING

Semi-supervised clustering, which uses class labels or pairwise constraints on some examples to aid unsupervised clustering, has been the focus of several recent projects (Basu, Banerjee, & Mooney, 2002; Klein, Kamvar, & Manning, 2002; Wagstaff, Cardie, Rogers, & Schroedl, 2001; Xing, Ng, Jordan, & Russell, 2003). If the initial labeled data represent all the relevant categories, then both semi-supervised clustering and semi-supervised classification algorithms can be used for categorization. However in many domains, knowledge of the relevant categories is incomplete.

III. LITERATURE SURVEY

A. ANALYSIS OF RELATED WORK

[1]. JOINT CLUSTER ANALYSIS OF ATTRIBUTE DATA AND RELATIONSHIP DATA: THE CONNECTED K-CENTER PROBLEM 2006

[2]. LEARNING DISTANCE METRICS WITH CONTEXTUAL CONSTRAINTS FOR IMAGE RETRIEVAL 2006

[3]. DISTANCE METRIC LEARNING, WITH APPLICATION TO CLUSTERING WITH SIDE-INFORMATION 2006

[4]. Maximum Margin Clustering Made Practical Proc.int'1 conf, (Machine Learning) k.Zhang, i.w.Tsang, and J.T.Kwok.2007

[5]. Learning Nonparametric Kernel Matrices from Pairwise Constraints Proc. Int'l Conf. Machine Learning,. S.C.H. Hoi, R. Jin, and M.R. Lyu2007

[6]. GENERALIZED MAXIMUM MARGIN CLUSTERING AND UNSUPERVISED KERNEL LEARNING Hamed Valizadegan, Rong Jin 2007

[7]. Transductive Support Vector Machines for Structured Variables Proc. Int'l Conf. MachineLearning, A. Zien, U. Brefeld, and T. Scheffer 2007

[8]. Pegasos: Primal Wstimated Sub-Gradient Solver for SVM Proc. Int'l Conf. MachineLearning, S. Shalev-Shwartz, Y. Singer, and N. Srebro2007

[9]. Constrained Spectral Clustering through Affinity Propagation Proc. IEEE Conf. Computer Vision and Pattern Recognition. Z. Lu and M.A. Carreira-Perpinan2008

[10]. Semi-supervised Clustering with Metric Learning Using Relative Comparisons Knowledge and Data Eng., N. Kumar and K. Kummamuru 2008

[11]. Efficient Multiclass Maximum Margin Clustering Proc. Int'l Conf. Machine Learning, B. Zhao, F. Wang, and C. Zhang,2008

[12]. Maximum Margin Clustering with Pairwise Constraints Proc. Int'l Conf. Data Mining Y. Hu, J. Wang, N. Yu, and X.S. Hua,2008

[13]. Improving Classification with Pairwise Constraints: A Margin-Based Approach Proc. European Conf. Machine Learning and Knowledge Discovery in Databases, N. Nguyen and R. Caruana2008

[14]. Large Scale Manifold Transduction, Proc. Int'l Conf. Machine Learning, M. Karlen, J. Weston, A. Erkan, and R. Collobert,.

[15]. Maximum Margin Clustering Made Practical,Neural Networks K. Zhang, I.W. Tsang, and J.T. Kwok, "2009

[16]. TIGHTER AND CONVEX MAXIMUM MARGIN CLUSTERING Yu-Feng Li1 Ivor W. Tsang2 James T. Kwok3 Zhi-Hua Zhou12009

[17]. CLUSTERING WITH MULTIVIEWPOINT-BASED SIMILARITY MEASURE Duc Thang Nguyen, Lihui Chen, Senior Member, IEEE, and Chee Keong Chan IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 24, NO. 6, JUNE 2012

[18]. SEMI-SUPERVISED MAXIMUM MARGIN CLUSTERING WITH PAIRWISE COSTRAINTS Hong Zeng , Member , IEEE,and Yiu-Ming Cheung, Senior Member,IEEE 2012

IV. EXISTING SYSTEM

In existing system approaches for clustering data are based on metric similarities, i.e., nonnegative, symmetric, and satisfying the triangle inequality measures. For text mining tasks, the majority of state-of-the-art frameworks employ the vector space model (VSM), which treats a document as a bag of words and uses plain language words as features. This model can represent the text mining problems easily and directly. However, with the increase of data set size, the vector space becomes high dimensional, sparse, and the computational complexity grows exponentially. On the other hand, supervised learning needs an initial large number of class label information, which requires expensive human labor and time.

A. SUPERVISED LEARNING

Supervised learning is the machine learning task of inferring a function from supervised training data. The training data consist of a set of training examples. In supervised learning, each example is a *pair* consisting of an input object (typically a vector) and a desired output value (also called the supervisory signal). A supervised learning algorithm analyzes the training data and produces an inferred function, which is called a classifier (if the output is discrete, see classification) or a regression function (if the output is continuous, see regression). The inferred function should predict the correct output value for any valid input object. This requires the learning algorithm to generalize from the training data to unseen situations in a "reasonable" way (see inductive bias). (Compare with unsupervised learning.) The parallel task in human and animal psychology is often referred to as concept learning.

V. PROPOSED SYSTEM

In our proposed system implements, semi supervised learning has captured a great deal of attentions. Semi supervised learning is a machine learning paradigm in which the model is constructed using both labeled and unlabeled data for training typically a small amount of labeled data and a large amount of unlabeled data. In this proposed system it retrieve the data from training data or labeled data and extract the feature of the data and compare with labeled data and unlabeled data to. It reduce the human work that need not train all data in the label data it occupy less memory this method user to make an accurate clustering. In clustering process, semi-supervised learning is a class of machine learning techniques that make use of both labeled and unlabeled data for training - typically a small amount of labeled data with a large amount of unlabeled data. Semi-supervised learning falls between unsupervised learning (without any labeled training data) and supervised learning (with completely labeled training data). Many machine-learning researchers have found that unlabeled data, when used in conjunction with a small amount of labeled data, can produce considerable improvement in learning accuracy.

Pairwise constrain clustering it contains the word comparison for more effective clustering. The semi supervised algorithm overcomes the untrained data process. Another drawback of the existing clustering is word pair process. Same meaning words come in different words example computer, computers, computing all belongs to the computer word. We compare these kind of data by using pair wise process on word match process based on the matching percentage we can assume that pair comes came category or different category. In existing system using K-mean the document will make cluster in over lapping or in different cluster that not relevant to any cluster groups. In this proposed system we overcome over lapping problems and avoid miss similar cluster problem too.

A. *ADVANTAGES*:

- Less training set and less memory will occupy by handling semi supervised process.
- Alike data can’t be miss in cluster data by using pairwise constrain.
- Overlapping avoid by using maximum margin cluster process.
- Apply correlation measure process

B. ALGORITHM

The k-means Algorithm

The k-means algorithm is an evolutionary algorithm that gains its name from its method of operation. The algorithm clusters observations into k groups, where k is provided as an input parameter. It then assigns each observation to clusters based upon the observation’s proximity to the mean of the cluster. The cluster’s mean is then recomputed and the process begins again. Here’s how the algorithm works:

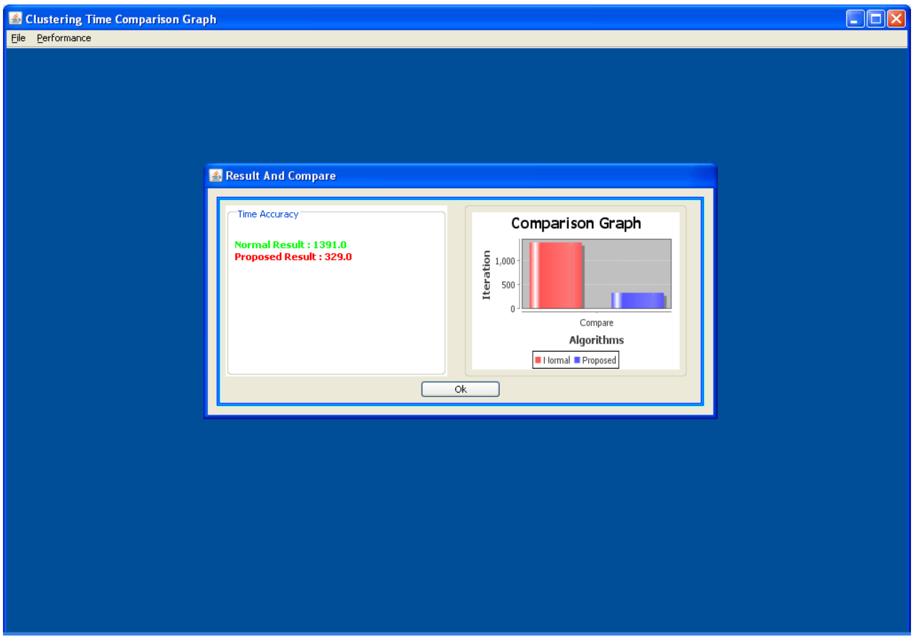
[1]. The algorithm arbitrarily selects k points as the initial cluster centers (“means”).
[2]. Each point in the dataset is assigned to the closed cluster, based upon the Euclidean distance between each point and each cluster center.
[3]. Each cluster center is recomputed as the average of the points in that cluster.
[4]. Steps 2 and 3 repeat until the clusters converge. Convergence may be defined differently depending upon the implementation, but it normally means that either no observations change clusters when steps 2 and 3 are repeated or that the changes do not make a material difference in the definition of the clusters.

C. PROPOSED ALGORITHM

Affinity Propagation

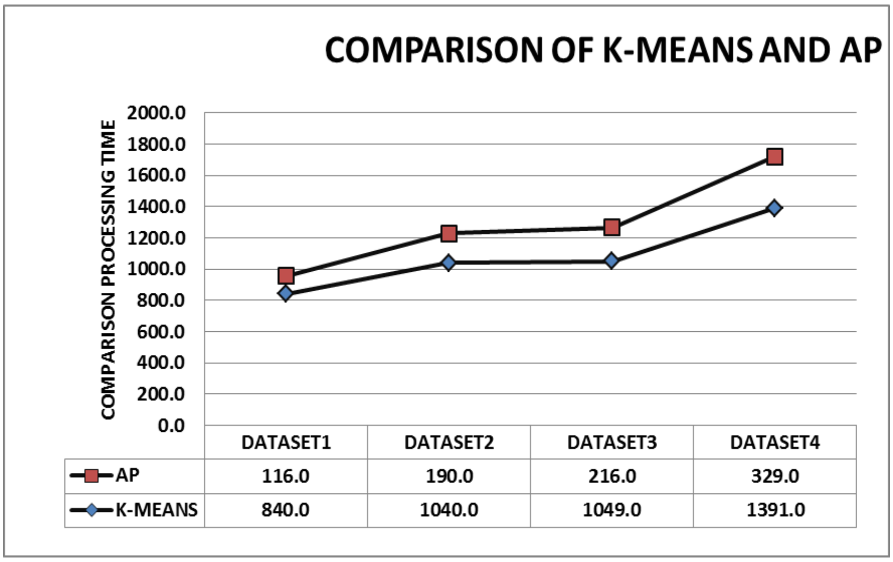
[1]. Clustering algorithm that works by finding a set of exemplars (prototypes) in the data and assigning other data points to the exemplars.
[2]. Input: pair-wise similarities (negative squared error), data point preferences (larger = more likely to be an exemplar)
[3]. Approximate maximization of the sum of similarities to exemplars
[4]. Semi-supervised Learning
[5]. Large amounts of unlabeled training data
[6]. Some limited amounts of side information
[7]. All points sharing the same label should be in the same cluster.
[8]. Points with different labels should not be in the same cluster

D. RESULT ANALYSIS



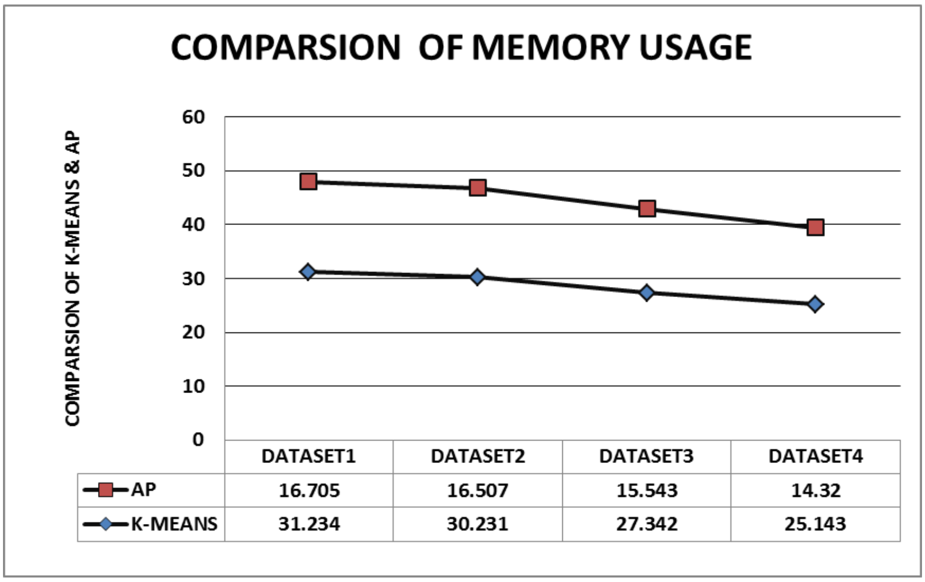
Normal Result : 1391.0
Proposed Result : 329.0

Fig. I Comparison of k-means and AP Time Accuracy



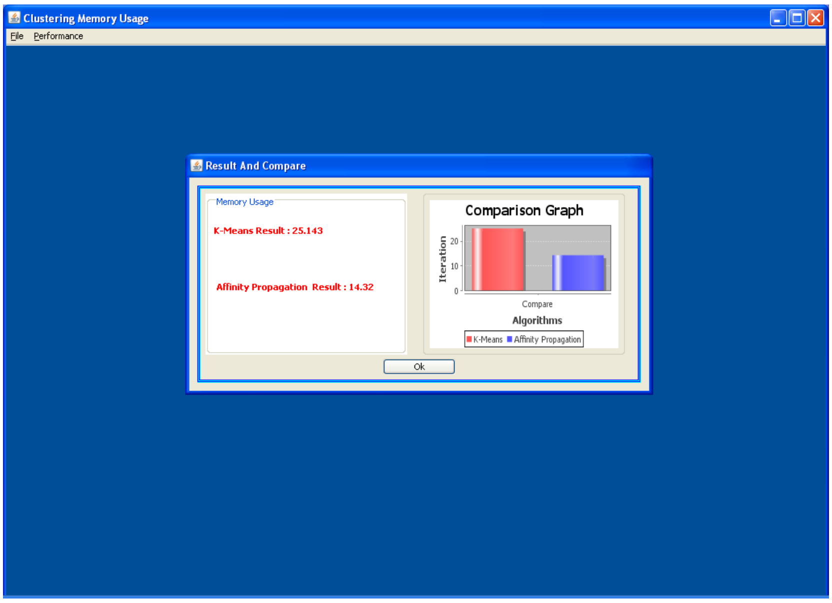
	DATASET1	DATASET2	DATASET3	DATASET4
AP	116.0	190.0	216.0	329.0
K-MEANS	840.0	1040.0	1049.0	1391.0

Fig. II Comparison of k-means and AP Time Accuracy



	DATASET1	DATASET2	DATASET3	DATASET4
AP	16.705	16.507	15.543	14.32
K-MEANS	31.234	30.231	27.342	25.143

Fig. III Comparison of k-means and AP Memory Usage



Memory Usage

K-Means Result : 25.143

Affinity Propagation Result : 14.32

Fig.IV Show Comparison of k-means and AP (Memory Usage)

From this table AP produce the best result in less time and in less memory space

TABLE I

S.No	Dataset Name	Average time (milli secs)		Memory usage (kb)	
		K-means	Ap	K-means	Ap
1	Dataset 1	840.0	116.0	31.234	16.705
2	Dataset2	1040.0	190.0	30.231	16.507
3	Dataset3	1049.0	216.0	27.342	15.543
4	Dataset4	1391.0	329.0	25.143	14.32

VI. CONCLUSION AND FUTURE WORK

A. Conclusion

Our main goal in the proposed thesis is to study search-based semi-supervised clustering algorithms and apply them to cluster the documents.

[1]. How supervision can be provided to clustering in the form of labeled data points or pairwise constraints;

[2]. How informative constraints can be selected in an active learning framework for the pairwise constrained semi-supervised clustering model; and

[3]. How search based and similarity-based techniques can be unified in semi-supervised clustering. In our work so far, we have mainly focused on generative clustering models, e.g. KMeans and EM, and ran experiments on clustering low-dimensional UCI datasets or high-dimensional text datasets.

In this thesis, we want to study other aspects of semi-supervised clustering, like:

[1]. The effect of noisy, probabilistic or incomplete supervision in clustering;

[2]. Model selection techniques for automatic selection of number of clusters in semi-supervised clustering;

[3]. Ensemble semi-supervised clustering. In future, we want to study the effect of semi-supervision on other clustering algorithms, especially in the discriminative clustering and online clustering framework. We also want to study the effectiveness of our semi-supervised clustering algorithms on other domains,

e.g., web search engines (clustering of search results), astronomy (clustering of Mars spectral images) and bioinformatics (clustering of gene microarray data).

B. Scope for Future Work

The future applies a novel clustering-based feature subset selection algorithm for high dimensional data. The algorithm involves

[1]. Removing irrelevant features,

[2]. Constructing a minimum spanning tree from relative ones, and

[3]. Partitioning the MST and selecting representative features. In the future algorithm, a cluster consists of features. Each cluster is treated as a single feature and thus dimensionality is drastically reduced.

Feature subset selection can be viewed as the process of identifying and removing as many irrelevant and redundant features as possible.

This is because

[1]. Irrelevant features do not contribute to the predictive accuracy, and

[2]. Redundant features do not redound to getting a better predictor for that they provide mostly information which is already present in other feature(s).

Traditionally, feature subset selection research has focused on searching for relevant features. A well-known example is Relief [34], which weighs each feature according to its ability to discriminate instances under different targets based on distance-based criteria function. However, Relief is ineffective at removing redundant features as two predictive but highly correlated features are likely both to be highly weighted [36]. Relief-F [37] extends Relief, enabling this method to work with noisy and incomplete data sets and to deal with multiclass problems, but still cannot identify redundant features.

REFERENCES

[1]. X. Wu, V. Kumar, J.R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G.J. McLachlan, A. Ng, B. Liu, P.S. Yu, Z.-H. Zhou, M. Steinbach, D.J. Hand, and D. Steinberg, “Top 10 Algorithms in Data Mining,” Knowledge Information Systems, vol. 14, no. 1, pp. 1-37, 2007.

[2]. Guyon, U.V. Luxburg, and R.C. Williamson, “Clustering: Science or Art?,” Proc. NIPS Workshop Clustering Theory, 2009.

[3]. Dhillon and D. Modha, “Concept Decompositions for Large Sparse Text Data Using Clustering,” Machine Learning, vol. 42, nos. 1/2, pp. 143-175, Jan. 2001.

[4]. S. Zhong, “Efficient Online Spherical K-means Clustering,” Proc. IEEE Int’l Joint Conf. Neural Networks (IJCNN), pp. 3180-3185, 2005.

[5]. Banerjee, S. Merugu, I. Dhillon, and J. Ghosh, “Clustering with Bregman Divergences,” J. Machine Learning Research, vol. 6, pp. 1705-1749, Oct. 2005.

[6]. Pekalska, A. Harol, R.P.W. Duin, B. Spillmann, and H. Bunke, “Non-Euclidean or Non-Metric Measures Can Be Informative,” Structural, Syntactic, and Statistical Pattern Recognition, vol. 4109, pp. 871-880, 2006.

[7]. M. Pelillo, “What Is a Cluster? Perspectives from Game Theory,” Proc. NIPS Workshop Clustering Theory, 2009.

[8]. Lee and J. Lee, “Dynamic Dissimilarity Measure for Support Based Clustering,” IEEE Trans. Knowledge and Data Eng., vol. 22, no. 6, pp. 900-905, June 2010.

[9]. Banerjee, I. Dhillon, J. Ghosh, and S. Sra, “Clustering on the Unit Hypersphere Using Von Mises-Fisher Distributions,” J. Machine Learning Research, vol. 6, pp. 1345-1382, Sept. 2005.

[10]. W. Xu, X. Liu, and Y. Gong, “Document Clustering Based on Non-Negative Matrix Factorization,” Proc. 26th Ann. Int’l ACM SIGIR Conf. Research and Development in Informaion Retrieval, pp. 267-273, 2003.

[11]. I.S. Dhillon, S. Mallela, and D.S. Modha, “Information-Theoretic Co-Clustering,” Proc. Ninth ACM SIGKDD Int’l Conf. Knowledge Discovery and Data Mining (KDD), pp. 89-98, 2003.

[12]. C.D. Manning, P. Raghavan, and H. Schu¨ tze, An Introduction to Information Retrieval. Cambridge Univ. Press, 2009.

[13]. Ding, X. He, H. Zha, M. Gu, and H. Simon, “A Min-Max Cut Algorithm for Graph Partitioning and Data Clustering,” Proc. IEEE Int’l Conf. Data Mining (ICDM), pp. 107-114, 2001.

[14]. H. Zha, X. He, C. Ding, H. Simon, and M. Gu, “Spectral Relaxation for K-Means Clustering,” Proc. Neural Info. Processing Systems (NIPS), pp. 1057-1064, 2001.

[15]. J. Shi and J. Malik, “Normalized Cuts and Image Segmentation,” IEEE Trans. Pattern Analysis Machine Intelligence, vol. 22, no. 8, pp. 888-905, Aug. 2000.

[16]. I.S. Dhillon, “Co-Clustering Documents and Words Using Bipartite Spectral Graph Partitioning,” Proc. Seventh ACM SIGKDD Int’l Conf. Knowledge Discovery and Data Mining (KDD), pp. 269-274, 2001.

[17]. Y. Gong and W. Xu, Machine Learning for Multimedia Content Analysis. Springer-Verlag, 2007.

[18]. Y. Zhao and G. Karypis, “Empirical and Theoretical Comparisons of Selected Criterion Functions for Document Clustering,” Machine Learning, vol. 55, no. 3, pp. 311-331, June 2004.

[19]. Karypis, “CLUTO a Clustering Toolkit,” technical report, Dept. of Computer Science, Univ. of Minnesota, http://glaros.dtc.umn. edu/~gkhome/views/cluto, 2003.

[20]. Strehl, J. Ghosh, and R. Mooney, “Impact of Similarity Measures on Web-Page Clustering,” Proc. 17th Nat’l Conf. Artificial Intelligence: Workshop of Artificial Intelligence for Web Search (AAAI), pp. 58-64, July 2000.

[21]. Ahmad and L. Dey, “A Method to Compute Distance Between Two Categorical Values of Same Attribute in Unsupervised Learning for Categorical Data Set,” Pattern Recognition Letters, vol. 28, no. 1, pp. 110-118, 2007.

[22]. Ienco, R.G. Pensa, and R. Meo, “Context-Based Distance Learning for Categorical Data Clustering,” Proc. Eighth Int’l Symp. Intelligent Data Analysis (IDA), pp. 83-94, 2009.

[23]. P. Lakkaraju, S. Gauch, and M. Speretta, “Document Similarity Based on Concept Tree Distance,” Proc. 19th ACM Conf. Hypertext and Hypermedia, pp. 127-132, 2008.

[24]. Chim and X. Deng, “Efficient Phrase-Based Document Similarity for Clustering,” IEEE Trans. Knowledge and Data Eng., vol. 20, no. 9, pp. 1217-1229, Sept. 2008.

[25]. S. Flesca, G. Manco, E. Masciari, L. Pontieri, and A. Pugliese, “Fast Detection of xml Structural Similarity,” IEEE Trans. Knowledge and Data Eng., vol. 17, no. 2, pp. 160-175, Feb. 2005.

[26]. E.-H. Han, D. Boley, M. Gini, R. Gross, K. Hastings, G. Karypis, V. Kumar, B. Mobasher, and J. Moore, “Webace: A Web Agent for Document Categorization and Exploration,” Proc. Second Int’l Conf. Autonomous Agents (AGENTS ’98), pp. 408-415, 1998.