Evolutionary Image Co-clustering with User Feedbacks

Amit Salunke[1], Manjeet Rege[2], and Reynold Bailey[3]

[1]C3 Energy, *amit.salunke@c3energy.com*
[2]University of St.Thomas, *rege@stthomas.edu*
[3]Rochester Institute of Technology, *rjb@cs.rit.edu*

Abstract

Traditional image clustering systems are primarily based on visual and/or textual features. Such algorithms commonly suffer from the problem of semantic gap. General approach to overcome this problem is to incorporate user feedback. However, data extracted from certain domains like social networks, web-blogs etc. is evolving in nature. In other words, data collected from such domains over small period of time interval exhibits high similarity over data instances and features, which can be effectively used to optimize data clustering since change in clustering is gradual. In this paper, we propose EHCC (Evolutionary star-structured Heterogeneous Co-Clustering) algorithm for image co-clustering. Our algorithm incorporates user provided feedbacks over period of time to guide co-clustering process. We incorporate user provided feedback in terms of image similarity logs over period of time to augment relational matrix obtained from low level features (color and textual features) extracted from images. Through an iterative algorithm, we tri-factorize new relational matrix to obtain image clusters. Through extensive experiments on image data sets, we demonstrate effectiveness and efficiency of our proposed algorithm.

keywords: image, clustering, user, feedback, matrix, factorization.

1 Introduction

Rapid development in data acquisition technology has resulted in the generation of large amounts of multimedia data. Consequently, problem of image clustering has gained lot of attention because of extensive applications in various domains like medical, criminal suspect tracking, social media, etc. There has been considerable work done on image clustering using low-level visual features. However, these approaches are unable to translate visual features to a sematic level leading to the problem of a semantic gap. To overcome this, relevance feedback mechanism was introduced to guide the clustering process. In a typical feedback mechanism, user marks few relevant images from the image dataset. The system then utilizes the semantics embedded in the feedback to guide the clustering process.

Images extracted from certain domains like multimedia [1], biomedicine [9] [18] [28], web mining [19] [26], etc. are evolving in nature. In such domains, data collected over small period of time interval exhibits high similarity over data instances and features. This can be effectively used to optimize data clustering since change in clustering is gradual [3]. For example, in Figure 1, initially lion and elephant images were clustered together as animals. However, over a period of time, as more images are collected, we have sufficient images to form separate clusters for images of lions and elephants. Accordingly, the user would begin to notice this difference and mark them in the feedback as well. None of the current proposed algorithms [4] [10] [5] take into account the passage of time and knowledge that can be gained by observing data as it evolves. Since, evolutionary clustering provides gradual change in data clustering over period of time, it is more tolerant to data noise. Chakrabarti et al. [3] and Chi et al. [7] have shown incorporation of historic knowledge improves the clustering accuracy. Wang et al. [25] have shown that evolutionary clustering can be achieved by combining low-rank matrix approximation methods and matrix factorization based clustering. Recently, Green et al. [12] proposed evolutionary spectral co-clustering approach for evolutionary data.

In this paper, we propose a novel algorithm designed for co-clustering of evolving image data with user feedback. We project co-clustering of evolutionary image data as an evolutionary star-structured heterogeneous co-clustering problem in which we perform non-negative matrix factorization over multiple time slices to handle evolving image data with user feedbacks [20] [5] [4] [10]. In order to perform co-clustering on evolutionary heterogeneous data, it is necessary to augment current data using historical data [3] [7] [12]. Then, we perform

978-1-880843-98-7 / copyright ISCA, CATA 2015
March 9-11, 2015, Honolulu, Hawaii, USA

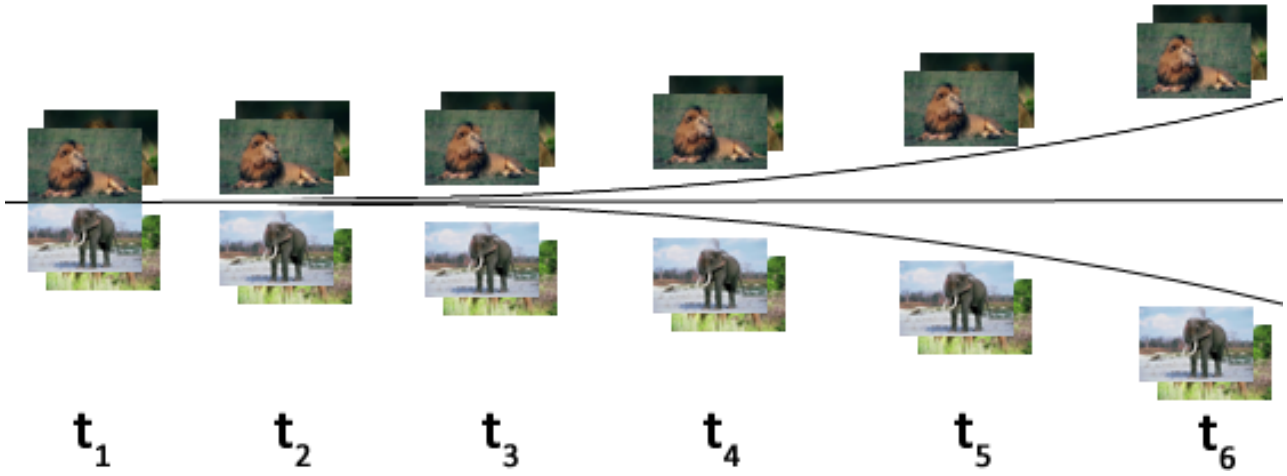

Figure 1: Visualization of Evolutionary image dataset, showing separation of two clusters, namely *African-lion* and *African elephant* images over period of time t_0 to t_6 which are initially clustered together as *African animals*

non-negative matrix factorization on the augmented current data to infer image clustering [6] [20] [11] [15].

2 Evolutionary Image Co-clustering with User Feedbacks

In the star-structured heterogeneous image data, at a time step t, we represent data using relational matrices with central image data type c connected to color-feature data, texture-feature data and image-log data as shown in Figure 2. We represent their relationship with the central data type using relational matrices $W_t^{(c1)}$, $W_t^{(c2)}$, and $W_t^{(c3)}$ respectively.

One of the challenges of evolutionary data is the dynamic nature of the data. Over the period of time, we may have a change in data size (change in number of samples and features of data) as well as change in the structure of data (change in number of clusters in the data). Low rank approximation methods extract correlation and then remove redundancy from data to obtain sparse data. This helps to make the original data mining algorithm computationally efficient on large data sets. To integrate historic data into current data, we make use of the family of Colibri methods [23]. First, Colibri-S iteratively constructs optimized sub-space by eliminating redundant columns. Colibri-D makes use of sub-space calculated at time t to calculate sub-space at time *t-1*. Since the change in graph between two consecutive time steps is assumed to be reasonably small, the overall edges effect between two time steps are far too less. Hence, for dynamic data, once we

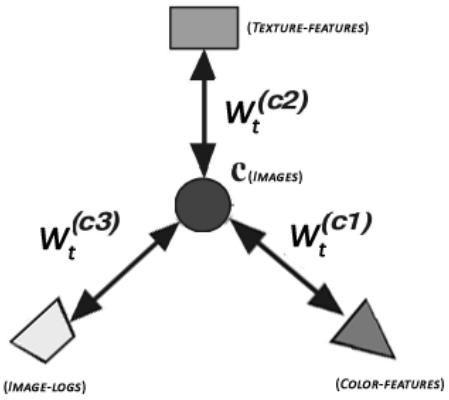

Figure 2: Visualization of Image data as Star-structured High-order Heterogeneous Data consisting texture-feature, color-features and user-logs connected to central data type, images

obtain initial sub-space using Colibri-S, we update sub-space for modified data over period of time. This is a faster method to update output matrix by sampling and comparing sampled edges between two time steps t and *t-1*. The output thus obtained is compatible with respect to current data and is computationally efficient. In the case, where historical data has less number of instances and/or features, we determine differences in instances and/or features between historical data and current data, and insert those many instances and/or features into historical data matrix. The inserted instance or feature receives the average of the whole matrix at current time step for each cell. So, we

apply low rank approximation to project original data $W_t^{(ci)} \in \Re^{n_c \times n_i}$ into subspace $\widetilde{W_t}^{(ci)} \in \Re^{n_c \times n_i}$ to improve computational efficiency of the algorithm in both time and space by hardly affecting the clustering accuracy.

Since, shape of data changes over the period of time, we estimate number of clusters using six different well known methods for cluster estimation. These methods are Silhouettes [21], Davies-Bouldin index [8], Calinski-Harabasz index [2], Krzanowski-Lai index [14], Hartigan [13], and weighted inter-to intra-cluster ratio with Homogeneity and Separation index [22]. The fact that the data is gradually evolving, we make use of cluster estimation on historical data to estimate number of clusters in current data. The process is faster since we need to estimate number of clusters closer to previously estimated value only. We estimate number of clusters using previously mentioned methods and use mode from above estimation methods to decide number of clusters in current data.

We define the overall cost function of EHCC (Evolutionary star-structured Heterogeneous Co-Clustering) as the sum of snapshot quality and historical cost. We solve this problem by maximizing the clustering quality of the current snapshot and minimizing the historical cost which provides clustering smoothness [3] [7] [25]. We propose the optimization equation for multiple time step data represented by, $\widetilde{W_t}^{(ci)} \in \Re^{n_c \times n_i}$ for $t=\{1, 2,..., S\}$ where is S is last time-step in data,

$$J = min_{L_{(t)}^{(c)}, M_{(t)}^{(ci)}, R_{(t)}^{(i)} \geq 0} \sum_{t=1}^{S} \alpha(1-\alpha)^{S-t} \cdot \| \widetilde{W_t}^{(ci)} - L_t^{(c)} M_t^{(ci)} R_t^{(i)} \|^2 \quad (1)$$

where α is trade off between historic data (historic quality) and current data (snapshot quality), and $0 \leq \alpha \leq 1$. $L_t^{(c)} \in \Re^{n_c \times k_c}$ (row coefficient matrix) and $R_t^{(i)} \in \Re^{k_i \times n_i}$ (column coefficient matrix) are indicator matrices representing soft-clustering for instances and features respectively and $M_t^{(ci)} \in \Re^{k_c \times k_i}$ (block value matrix) is co-relation indicator matrix that represents co-clustering relation between central data type and each data type connected with central data type at time step t [16] [27] [6]. We solve the above optimization problem using an iterative solution to obtain $L_S^{(c)}$, $R_S^{(i)}$ and $M_S^{(ci)}$ [25] given by,

$$L_{(S)(ab)}^{(c)} \leftarrow L_{(S)(ab)}^{(c)} \cdot \frac{\sum_{i=1}^{p} ((\sum_{t=1}^{S} \alpha(1-\alpha)^{S-t} \cdot \widetilde{W_t^{(ci)}}) R_{(S)}^{(i)\,T} M_{(S)}^{(ci)\,T})_{ab}}{\sum_{i=1}^{p} (L_{(S)}^{(c)} M_{(S)}^{(ci)} R_{(S)}^{(i)} R_{(S)}^{(i)\,T} M_{(S)}^{(ci)\,T})_{ab}} \quad (2)$$

$$R_{(S)(ab)}^{(i)} \leftarrow R_{(S)(ab)}^{(i)} \cdot \frac{(L_{(S)}^{(c)\,T} (\sum_{t=1}^{S} \alpha(1-\alpha)^{S-t} \cdot \widetilde{W_t^{(ci)}}) R_{(S)}^{(i)\,T})_{ab}}{(L_{(S)}^{(c)\,T} L_{(S)}^{(c)} M_{(S)}^{(ci)} R_{(S)}^{(i)} R_{(S)}^{(i)\,T})_{ab}} \quad (3)$$

$$M_{(S)(ab)}^{(ci)} \leftarrow M_{(S)(ab)}^{(ci)} \cdot \frac{(L_{(S)}^{(c)\,T} (\sum_{t=1}^{S} \alpha(1-\alpha)^{S-t} \cdot \widetilde{W_t^{(ci)}}) R_{(S)}^{(i)\,T})_{ab}}{(L_{(S)}^{(c)\,T} L_{(S)}^{(c)} M_{(S)}^{(ci)} R_{(S)}^{(i)} R_{(S)}^{(i)\,T})_{ab}} \quad (4)$$

3 Experiments and Results

The image data used in our experiments is chosen from Corel-CDs, which contains 31,438 general-purpose images from different catagories like animal, automobiles, hairstyles, waterfall, landscape, etc. For image co-clustering, we represented each image in the form of vector of 45 color features, 42 texture features [17] [24]. The color features include color channels (RGB, 9 features, including mean, variance, and skewness of R, G, and B channels), color histogram (CH, 12 features), and color coherence vector (CCV, 24 features). Texture features include Gabor wavelet based texture (Gab, 24 features), edge direction histogram (EDH, 9 features), and edge direction coherence vector (EDCV, 9 features). We constructed two matrices, *image-color* and *image-texture* representing color and textures features respectively.

We selected 500 images from 5 different contents, namely, *African Lion*, *African Elephant*, *Sunsets*, *Bonsai*, and *Aviation*. Some examples from each category are shown in Figure 3. In the proposed relevance feedback framework, we collect the users' positive feedback as samples to construct image-log matrix. Through feedback, the images marked indicate that they are similar to each other according to user's preference also called as a log. In the *image-log* matrix, for every log, user marks 3-5 similar images from 20 images randomly selected from the image pool. Images that are similar in a given log are marked 1, while the rest of the entries for that log would be a *zero*.

We increase the number of logs per category over the period of time t_0 to t_{10}. We mark all clusters distinctly with logs, so over a period of time, we expect more tight intra-cluster association, and different clusters are more distinctly identified. Then, we compare results for evolutionary algorithm with two different values of trade off factor α, which equal to 0.8 and 0.2. This experiment also highlights algorithms ability of historic knowledge.

Figure 3: Image samples selected random from each image category

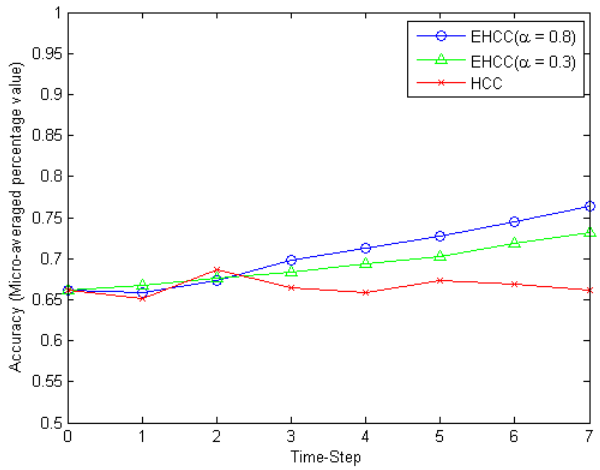

Figure 4: Comparison between evolutionary higher order clustering algorithm (EHCC) and Static Higher order clustering algorithm (HCC), and comparison of performances of evolutionary algorithm with respect to trade-of factor α

In Figure 4, we compare evolutionary with static version of higher order co-clustering. As expected, evolutionary algorithm produces more accurate results than the static version of higher-order co-clustering algorithm due to its ability to integrate historic knowledge. When we lowered the value of trade-off factor, i.e. $\alpha = 0.3$, we gave more importance to historical data, leading to a lower clustering accuracy as expected.

Also, to evaluate whether the algorithm is able to handle cluster evolution over period of time, initially we marked *African Lion* and *African Elephant* to have them belong to the same cluster. Over a period of time t_0 to t_7, we marked them separately into two clusters. The algorithm was successfully able to learn from the evolving user feedback, and was able to group the lions and the elephants into two different clusters eventually.

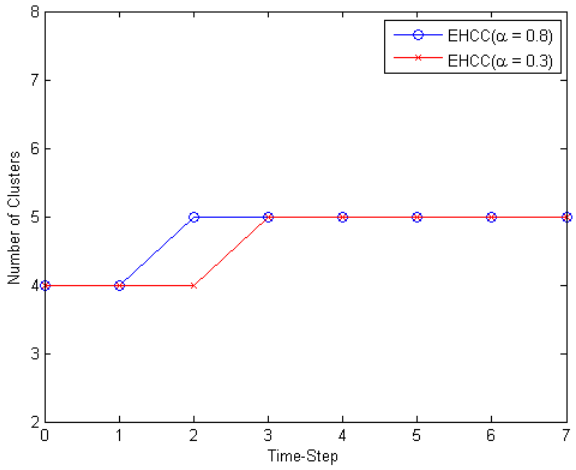

Figure 5: Change in number of clusters over period of time, for EHCC and HCC; and comparison of performances of evolutionary algorithm with respect to trade-of factor α

From Figure 5, we show that the algorithm is able to adapt well to change in data. A new cluster, i.e. 5^{th} cluster, is discovered at time-step t_2, even though we have started marking 5^{th} from time-step t_1. So, a new cluster is not formed until we have enough number of instances in the new cluster for a sufficient period of time. For trade-off value $\alpha = 0.3$, we gave more importance for historical data, cluster change occurred more gradually. Hence, the fifth image cluster was detected one time-step later at t_3. In Figure 6, we show the number of instances clustered into the 5^{th} cluster. We can observe that over a period time, the 5^{th} cluster becomes more prominent. As shown earlier, when we lower the trade-off value, as a result of giving more importance to hisorical data, the algorithm is more focused on the past, and is unable to adapt to the change.

4 Conclusions

In this paper, we present evolutionary higher-order co-clustering algorithm for image clustering. We integrate user feedback in terms of similarity logs to perform higher-order co-clustering on low level features extracted from images. We integrate user feedback over period of time to augment current data matrix, then perform higher order non-negative matrix factor-

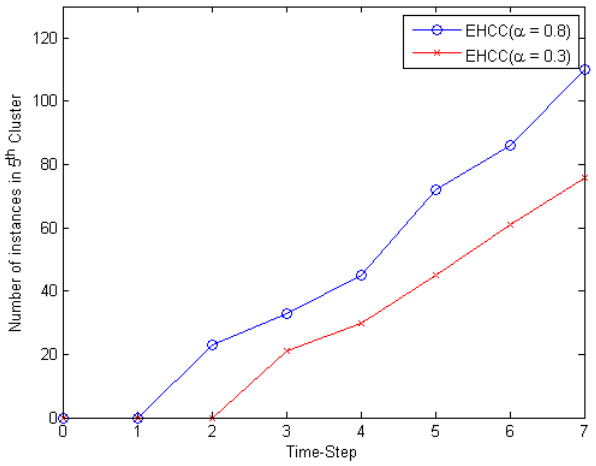

Figure 6: Number of instances in 5^{th} cluster over period of time, for EHCC and HCC; and comparison of performances of evolutionary algorithm with respect to trade-of factor α

ization to obtain co-clustering at current time-step. Empirically, we show effectiveness of our algorithm for handling evolutionary data, and its stability in clustering than a static approach.

References

[1] Rui Cai, Lie Lu, and Alan Hanjalic. Unsupervised content discovery in composite audio. In *Proceedings of the 13th annual ACM international conference on Multimedia*, MULTIMEDIA '05, pages 628–637, New York, NY, USA, 2005. ACM.

[2] T. Calinski and J. Harabasz. A dendrite method for cluster analysis. *Communications in Statistics*, 3:1–27, 1974.

[3] Deepayan Chakrabarti, Ravi Kumar, and Andrew Tomkins. Evolutionary clustering. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '06, pages 554–560, New York, NY, USA, 2006. ACM.

[4] Yanhua Chen, Ming Dong, and Wanggen Wan. Image co-clustering with multi-modality features and user feedbacks. In *Proceedings of the 17th ACM international conference on Multimedia*, MM '09, pages 689–692, New York, NY, USA, 2009. ACM.

[5] Yanhua Chen, Manjeet Rege, Ming Dong, and Farshad Fotouhi. Deriving semantics for image clustering from accumulated user feedbacks. In *Proceedings of the 15th international conference on Multimedia*, MULTIMEDIA '07, pages 313–316, New York, NY, USA, 2007. ACM.

[6] Yanhua Chen, Lijun Wang, and Ming Dong. Non-negative matrix factorization for semisupervised heterogeneous data coclustering. *Knowledge and Data Engineering, IEEE Transactions on*, 22(10):1459 –1474, oct. 2010.

[7] Yun Chi, Xiaodan Song, Dengyong Zhou, Koji Hino, and Belle L. Tseng. Evolutionary spectral clustering by incorporating temporal smoothness. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '07, pages 153–162, New York, NY, USA, 2007. ACM.

[8] David L. Davies and Donald W. Bouldin. A cluster separation measure. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, PAMI-1(2):224 –227, april 1979.

[9] Chris Ding, Xiaofeng He, Richard F. Meraz, and Stephen R. Holbrook. A unified representation for multi-protein complex data for modeling protein interaction networks. *Proteins: Structure, Function, and Bioinformatics*, 57:99–108, 2004.

[10] Bin Gao, Tie-Yan Liu, Tao Qin, Xin Zheng, Qian-Sheng Cheng, and Wei-Ying Ma. Web image clustering by consistent utilization of visual features and surrounding texts. In *Proceedings of the 13th annual ACM international conference on Multimedia*, MULTIMEDIA '05, pages 112–121, New York, NY, USA, 2005. ACM.

[11] Bin Gao, Tie-Yan Liu, Xin Zheng, Qian-Sheng Cheng, and Wei-Ying Ma. Consistent bipartite graph co-partitioning for star-structured high-order heterogeneous data co-clustering. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, KDD '05, pages 41–50, New York, NY, USA, 2005. ACM.

[12] Nathan Green, Manjeet Rege, Xumin Liu, and Reynold Bailey. Evolutionary spectral co-clustering. In *Neural Networks (IJCNN), The 2011 International Joint Conference on*, pages 1074 – 1081, 31 2011-aug. 5 2011.

[13] John A. Hartigan. *Clustering Algorithms*. Wiley, New York, 1975.

[14] W J Krzanowski and Y T Lai. A criterion for determining the number of groups in a data

set using sum-of-squares clustering. *Biometrics*, 44(1):23–34, 1988.

[15] Bo Long, Zhongfei (Mark) Zhang, Xiaoyun Wú, and Philip S. Yu. Spectral clustering for multi-type relational data. In *Proceedings of the 23rd international conference on Machine learning*, ICML '06, pages 585–592, New York, NY, USA, 2006. ACM.

[16] Bo Long, Zhongfei (Mark) Zhang, and Philip S. Yu. Co-clustering by block value decomposition. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, KDD '05, pages 635–640, New York, NY, USA, 2005. ACM.

[17] Wei-Ying Ma and Hong Jiang Zhang. Benchmarking of image features for content-based retrieval. In *Conference Record of the Thirty-Second Asilomar Conference IEEE on Signals, Systems & Computers, 1998*, volume 1, pages 253–257, 1998.

[18] Ruggero G. Pensa and et al. Constrained co-clustering of gene expression data, 2008.

[19] M. Rege, M. Dong, and F. Fotouhi. Co-clustering image features and semantic concepts. In *Image Processing, 2006 IEEE International Conference on*, pages 137 –140, oct. 2006.

[20] Manjeet Rege, Ming Dong, and Jing Hua. Clustering web images with multi-modal features. In *Proceedings of the 15th international conference on Multimedia*, MULTIMEDIA '07, pages 317–320, New York, NY, USA, 2007. ACM.

[21] Peter Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.*, 20:53–65, November 1987.

[22] K. Tasdemir and E. Merenyi. A new cluster validity index for prototype based clustering algorithms based on inter- and intra-cluster density. In *Neural Networks, 2007. IJCNN 2007. International Joint Conference on*, pages 2205 – 2211, aug. 2007.

[23] Hanghang Tong, Spiros Papadimitriou, Jimeng Sun, Philip S. Yu, and Christos Faloutsos. Colibri: fast mining of large static and dynamic graphs. In *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '08, pages 686–694, New York, NY, USA, 2008. ACM.

[24] A. Vailaya, A. Jain, and H. J. Zhang. On image classification: City vs. landscape. In *Proceedings of the IEEE Workshop on Content - Based Access of Image and Video Libraries*, CBAIVL '98, pages 3–, Washington, DC, USA, 1998. IEEE Computer Society.

[25] Lijun Wang, Manjeet Rege, Ming Dong, and Yongsheng Ding. Low-rank kernel matrix factorization for large scale evolutionary clustering. *IEEE Transactions on Knowledge and Data Engineering*, 99(PrePrints), 2010.

[26] Guandong Xu, Yu Zong, Peter Dolog, and Yanchun Zhang. Co-clustering analysis of weblogs using bipartite spectral projection approach. In *Proceedings of the 14th international conference on Knowledge-based and intelligent information and engineering systems: Part III*, KES'10, pages 398–407, Berlin, Heidelberg, 2010. Springer-Verlag.

[27] Wei Xu, Xin Liu, and Yihong Gong. Document clustering based on non-negative matrix factorization. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, SIGIR '03, pages 267–273, New York, NY, USA, 2003. ACM.

[28] Sungroh Yoon, Luca Benini, and Giovanni De Micheli. Co-clustering: A versatile tool for data analysis in biomedical informatics. *IEEE Transactions on Information Technology in Biomedicine*, 11(4):493–494, 2007.