IOSR Journal Of Applied Physics (IOSR-JAP)
e-ISSN: 2278-4861.Volume 12, Issue 1 Ser. III (Jan. – Feb. 2020), PP 76-82
www.Iosrjournals.Org

A Generative Model for Discovering Photo Communities in Personal Albums

Dr. Vandna Bhalla
Associate Professor, Department of Electronics, Sri Aurobindo College, University of Delhi
Email: vbhalla.du@gmail.com

Abstract: *Overwhelming popularity and availability of inexpensive memory cards, high resolution yet cheap phone cameras and digital cameras has culminated in large amount of personal data occupying our storage spaces on devices and/or the web. The personal data have their own stories that cannot be captured by the low-level features. Content based image retrieval techniques are not adequate to bridge the so-called semantic gap between the content and the semantic interpretation of that information by a particular user. We present the study and implementation of a novel model which automatically identifies a particular individual's photo clusters on the basis of photo profile information and the arrangement of a graphical structure networking the various photos.*

Keywords: *Cluttered, Circles, Communities, Generative, Hierarchical, Membership, Metric, Nested, Photos, Profile, Similarity, Synthetic, Unsupervised,*

I. Introduction

Today's user wants to browse efficiently through huge personal collections made possible due to the easy and economical availability of digital cameras and immense storage at lower prices. The proliferation of low-cost multimedia devices has resulted in an unprecedented growth of personal data which may or may not be interrelated, is heterogeneous and unstructured. Full use and benefits of such collections are meaningless if the retrieval and access methods are limited and ineffective. It is labour as well as time intensive to select specific data for sharing or for personal use given the sheer size of these collections.

The personal photo collections for an individual in today's environment is an important category of personal data and it is cluttered, diverse and huge. These are collections of souvenirs of important events and contain precious memories in our lives. An individual stores these personal data to time and again refresh such instances by retrieving them. An individual also likes to share these given the many social networks that are available today. Many social networking platforms like for instance Twitter, Face book or Google+ have a facility to group contacts using the concept of social circles but it is a laborious manual task which requires recurrent user intervention every time the personal network is updated with additions or deletions in the contact lists. We present the study and implementation of a novel model which automatically identifies a particular individual's photo clusters on the basis of photo profile information and the arrangement of a graphical structure networking the various photos. In fact, management of these collections of personal digital photos is a big challenge and all of us have millions of these on our memory cards, web, or other storage devices. Unfortunately, there is no optimal technique available to automatically organize them systematically in a completely unsupervised manner. The model is based on clustering photos on the basis of a similarity metric which is specific to a particular photo cluster. Our experimental results are encouraging on diverse personal photo collections. The comparisons have been meticulously carried out with hand labelled ground truth which was an onerous time-consuming task but the results are comparable with the state of the art.

II. Challenges with Current Techniques

A multitude of prevailing algorithms and techniques result in good classification accuracy, but they require large training data, labelling and/or annotations. These are generally not very effective especially in the domain of personal data primarily due to limited set of labelled data for the many possible classes. It may happen that for some cases only limited labelled data is available but large unlabelled data is accessible. Most of the approaches to help organize personal data are primarily built on text-based search. This was a feasible mechanism when the amount of data and the number of users was small. These keyword-based mechanisms restrict intuitive content level access and fail to exploit the non-linear high level semantic relationships. Moreover, the lack of sufficient labelled data for training is an added challenge. The personalized systems built so far are mostly based on relevance feedback which requires explicit user intervention.

The on-line social network personal data also presents novel additional challenges. How to incorporate the vast and many a times unlinked information into personalized retrieval for different users. A user may desire to share his conference pictures with his colleagues at work and his birthday pictures with close family only. They may be inclined to limit certain content from their school friends. Evaluation is another challenge in retrieval which is context sensitive. Evaluation based on conventional judgement is not suited ideally to assess the extent to which context-based systems enhance the user's experience of information retrieval in practice. We realized that the following points are very unique and relevant in the context of personal data.

- The personal data is inherently very diverse.
- The inter and intra class diversity cannot be captured with normal standard retrieval procedures but require astute and intelligent techniques.
- Annotating all personal data is slow and boring often requiring expert knowledge and expensive and special devices.
- The individual collections are large and cluttered, but structures are unknown and the dynamics are transient.

Such functionality requires efficient techniques which are automatically updating, not time consuming and capturing the individual perspectives of their communities. New breakthroughs in personalized information retrieval can happen in two ways:

- More intelligent mathematical formulations capable of exploiting the prevailing information sources better, or
- New, fresh and novel algorithms for retrieval that can combine and use additional advanced sources of contextual meta data.

The proposed model is based on clustering photos on the basis of a similarity metric which is specific to a particular photo cluster. Our experimental results are encouraging on diverse personal photo collections. The comparisons have been meticulously carried out with hand labelled ground truth, which was an onerous time-consuming task, but the results are comparable with the state of the art.

III. Community Detection

We are now past the era of dealing with challenges related to data collection. Today the research is around management, storage, easy and quick retrieval for personal data of individuals. Clustering or Community detection is one such analytical technique where the data is segregated into groups such that data within the group is more similar with each other than with data in other groups /clusters [1]. It determines the intrinsic cluster in a set of unlabelled data. Image clustering today is an important area of computer vision research [2]. Data Clustering or grouping is an established technique in fields of computer science and finds wide applications in the areas of data mining, image processing, pattern recognition and artificial intelligence. Hierarchical and partition are two main approaches to clustering [3]. Hierarchical clustering techniques try to build a hierarchy of clusters and are usually represented by a tree called dendrogram. The Agglomerative or the bottom-up approach starts with each data in its own cluster and subsequently the cluster pairs are merged. All data points initially in one cluster is the top down or the Divisive approach. Recursively the splits are accomplished while moving down the hierarchy. Both the splits and the mergers are greedy in character. Generally, the complexity of the first type and the exhaustive search associated with the second type make both variants unsuitable. SOM, the Self-Organization Map algorithm is an ANN with unsupervised training and learning through a cooperation mechanism to accomplish dimensionality reduction. The topological attributes of the data are preserved using a neighbourhood function and competitive learning. Training mode uses the input sample to build the map and Mapping helps classify the new vectors automatically. Varied combining norms and distance measures can be experimented with to form clusters. The distance-based clustering algorithm, EM clustering technique, is popularly used in statistics. It extends the simple K-means approach. This algorithm finds the probabilities based on Probability Distributions of the cluster memberships, and computes classification probabilities. Given the clusters, the algorithm maximises the likelihood or the overall probability of the data. This can be applied to categorical as well as continuous variables.

Density based Clustering Algorithms groups closely packed points, that is having many neighbours close by, together. The techniques connect data points into a group if a density function is satisfied within permissible thresholds for distance. These invariably result in whimsical groupings but are noise resistant [4]. For almost all these techniques we need to assume the number of clusters in advance which is not practical for personal data albuming. Generally, as the number of clusters increase the performance starts to deteriorate, though Hierarchical improves when the number of clusters increase. K means, EM and Density based techniques are better than the Hierarchical Algorithm, but EM and K-means have low accuracy. The performance of EM and K-means are better with large datasets. Density based techniques are unable to handle high dimensionality. Based on the parameters of Dataset size, Performance, Dataset type, Quality, Time Complexity and Accuracy, there is no one clustering technique that does justice to all. The unsupervised learning algorithms try to capture a hidden invisible structure from the data but the accuracy evaluation in this domain is a challenge. Clustering, Anomaly detection, Neural networks are among some of the approaches used for Unsupervised learning. The classical algorithms in the past mostly cluster on the basis of graph structure [5] and [6]. The works of [7],[8] and [9] have shown that real world communities can overlap, and these were very significant inspirations for our work. Schaefer [10] talks about the disjoint communities and Ravasz and Barabasi [11] deal with hierarchical arrangements in real world communities. The work of Chang [12], Elka and Menon[13] and Vu et al[14] have put forth models based on probabilities of edges between two nodes without getting into the community formation concepts. They have presented similar models where two nodes form an edge based on probability but they do not culminate to form cluster/ circles/ communities based on similarity parameters. Yoshida in 2010[15] worked with social network data and performed clustering in that domain and so did J. McAuley and J. Leskovec in 2014[16]. Memberships to multiple communities was also taken into consideration. Yang and Leskovecs work in 2012[8] is also quite similar. It models communities of multiple overlap with hard memberships of member nodes based not on the features of the node but from the information delivered by the network. Communities have been identified using classical algorithms using graph structure [5] or features [17] but rarely with both in conjunction. Mixed memberships of entities to multiple clusters have been discovered by topic modelling techniques [18]. Text information attributes for entities are permitted by extensions [19].

IV. Model Proposed

The relevance and significance of personal data for the user is undergoing perpetual changes. Despite variations amongst the personal data collections like photos/videos, these assorted accumulations possess an associated context. Due to the diversity and the sheer size of data that will eventually build in a personal photo collection, an adaptive and personalized approach is required for building automated and organized albums to be able to provide better user experience and bridge the semantic gap created by the not so obvious relationship between the low-level features and the high-level concepts.

A generative model to discover communities in personal datasets is proposed to ease sharing and organization. We propose this as a clustering task where every node can have multi memberships. The novel model automatically discovers user's "photo communities" where each cluster is formed on the basis of some common photo similarity parameter and is a subset of the photo collection. The model detects groups of clusters to which a photo belongs by combining network structure as well as photo profile information i.e finds communities in photo collections. Clusters of similar photos are formed on the basis of common metrics between the photos. Personal photos are very distinct from other images as more than being just associated with an individual they have an explicit context that homogenizes them. The user typically will have associated recollections specifically relating the photographs like for instance the time, the place, the event, the people etc. A personal collection of a User may have photos of trips made to various countries on vacation at different times. Some of these countries may lie in the same continent. Pictures with various friends and/or family, wedding functions, picnics, pictures taken at same time every year etc. are some example clusters that maybe significant to every user. Figure 1 depicts some example clusters shown as circles from an individual's personal photo collections connected as a network.

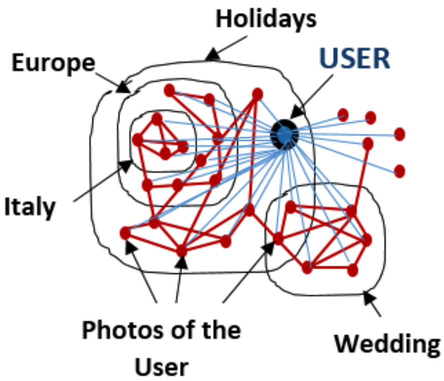

Figure 1: A personal photo network with labelled clusters

The User, U, is the owner of all the photos in his album and is labelled USER. A relational network is formed between all his pictures. We undertake the innovative task of identifying the particular clusters to which each photo will belong. The objective is to determine the photo communities or in other words photo collections based on similarities within a photo album. Following points are evident from Figure 1 and the model has been designed by taking these into cognizance.

- Photos within a cluster/circle/community should be similar under the particular parameter. Each circle will be formed under a specific aspect of the photo profile information.
- Different clusters will result from different similarity metrics. Model will be comprehensible by the user on the basis of the particular aspect of the similarity index that resulted in the formation of the cluster/circle/community.
- Communities will overlay i.e photos may simultaneously belong to many clusters at the same time (photos of a wedding held at a holiday destination X will overlap with photos of destination X).
- Hierarchical Nested Circles may be very common (photos of Italy and France will be nested inside the circle formed by photos shot in Europe).
- The photos may belong to none, one or many communities.

We propose a framework to learn photo communities and the profile similarity dimensions that created these densely connected photo clusters in a personal collection. The memberships of photos to clusters are modelled as latent variables. The similarity metrics between photos in a cluster appears as a function of the photo profile data. The completely unsupervised method helps to ascertain the specific parameter of photo sameness that culminated in the clusters.

V. Our Architecture

We propose an unsupervised model to find communities of photos from personal collections automatically. Our objective is to find user specific clusters of photos and to identify the common metric around which the photos group into the particular cluster. The photo communities have densely connected combinations of photos from the personal album sharing common attributes. The photos are permitted multiple memberships as they may belong simultaneously to multiple clusters. The circle affiliations are modelled as latent variables and the similarity between the photos is taken as a function of common parameters. Each photo cluster is permitted its own different profile parameter definition such that a cluster may form from photos of friends from the same institution and another one from photos of friends from the same area. The learning happens by choosing photo memberships and the similarity parameter simultaneously. We have evaluated our model on datasets made from real authentic photos from a personal collection. We also laboriously worked to establish the hand labelled ground truth for comparisons to ascertain the efficacy of our model. The experimental results show that the method has great potential for clustering personal photos in an unsupervised environment. Not much work has been reported in the literature for unsupervised personal photo clustering. The model in addition to being reasonably accurate determines the clusters as well as the number of clusters axiomatically. The methodology instinctively also explains the memberships of certain photos to common clusters. It is a significantly superior option compared to doing such cumbersome segregation by natural means. We propose a framework with the following features.

1. Photos within a community are related by a common profile metric.
2. The cluster/community/circles are identified by both the structure of the network of photos as well as their profiles.
3. Different clusters are allowed to form according to different similarity metrics emerging within the photo collections. Photos taken in the same year, for instance, will be grouped together and photos shot in a particular city will form another separate cluster. Likewise, the photos of a wedding or a picnic or a conference can be grouped separately. The attributes of the user/owner are not taken into consideration while grouping the personal photo collections. We build the model by connecting only the photo profile data.
4. Weaker circles lie within the stronger ones. Community of photos of Italy will lie within the circle having photos of Europe.
5. We present mixed membership of hard assignment models to multiple clusters but for a single user.
6. We have evaluated the model on two novel and original datasets. The first set is based on real pictures from a private user‘s collection. The second set is machine generated synthetic data which simulates real data. Since we are dealing with personal collections, the data size as well as the dimension size of data is limited.

We define a model for discovering photo communities given a network N of photos p2 P and a set of clusters O. Let E is the set of all edges between the photos in N. Such a network forms clusters based on any common metric between the photos. For each cluster its member photos are identified, and the cluster specific photo profile similarity metric is learnt. A set of clusters O = [O1;O2; :::::Ok];Ok P are formed based on a similarity parameter Sk which governs the formation of each cluster or circle. The photo profiles are expressed in such a way that they capture the common properties of any two photos. For instance a similarity dimension Sk of r(m,n) is important for circle Ok, where r(m,n) are pairwise feature vectors between two photos m and n. Each photo has a binary membership with a cluster and can also belong to more than one cluster if it satisfies the specific similarity criteria. The set of nodes in the graph represent all the photo members (pi). The set of edges, E, in the graph represent relationship between the photos. Input is a Photo Network, N = (P, E), along with profiles for each photo p 2P. Figure 2 illustrates the formulation of profile information for the dataset.

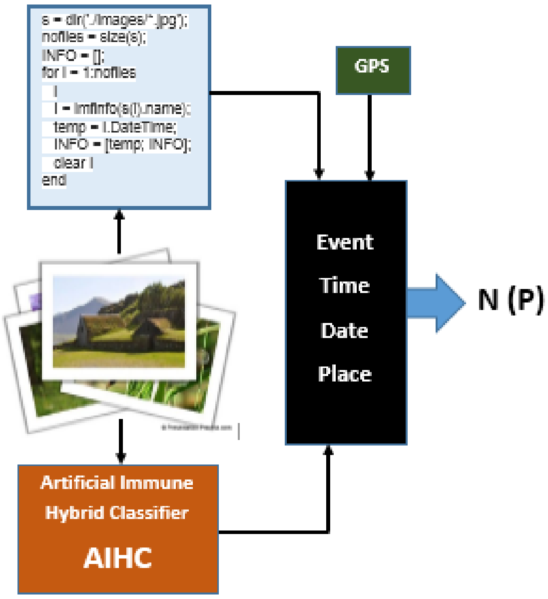

Figure 2: Creating N(P, E) using Profile Information

The Artificial Immune Hybrid Classifier (AIHC) [20], designed and implemented by us is used to identify the EVENTS in the personal photo dataset. The EXIF data is extricated using the small software to identify the YEAR and MONTH and the GPS data are used to locate the CONTINENT and COUNTRY where the photo was shot. More profile information can be added if desired by a user.

VI. Results

We have used two kinds of datasets for our experiments :(1) lab generated synthetic data and (2) pictures from personal photo albums. For the synthetic dataset we took upto 1000 samples each with 20 profiles. The photo album had a maximum of 149 samples each with 24 profiles. This method is completely unsupervised but to evaluate the performance, a fully labelled data for ground truth evaluation was built in the lab for both sets of data. This was a laborious time consuming task using considerable amount of resources, efforts and time. The number of clusters are optimally chosen automatically such that the number minimizes the Schwarz or the Bayesian Information Criteria (BIC) based on the likelihood function.

S.No	Profiles	Samples	Edges	Circles	SeedE	SeedC	F_1	J
1	10	100	9234	348	100	23	0.42	0.41
2	10	1000	9,39,694	1023	10,000	100	0.34	0.34
3	20	200	39,620	7000	7000	50	0.20	0.39

Table 1

We have conducted experiments for varying sample size and feature size respectively. Both the F1 score and J scores are reported. Table 1 tabulates results for synthetic data. The average F1 Score is 0.32 which deteriorates as the number of profiles increase. The average J score is 0.38 which remain steady with increase in profile dimensions.

S.No	Profiles	Samples	Edges	Circles	SeedE	SeedC	F_1	J
1	7	46	54	3	20	2	0.45	0.445
2	7	46	54	6	20	2	0.38	0.410
3	7	46	54	10	20	2	0.36	0.409

Table 2

Table 2 shows the result of experiments conducted on real photos. It shows a marked increase in accuracy. The synthetic data was completely random with no real connection like the ones existing in real world data. Hence the result deteriorated.

VII. Discussions and Conclusions

We have developed a model for similar photo community detection which can be used for each user independently. The model naturally learns the dimensions that generate clusters/circles /communities. The communities detected are very close to the ground truth which have been manually labelled. The method is able to identify overlapping as well as nested clusters. The model automatically learns the metric that leads to the formation of the particular cluster. This model can be applied to different users independently to generate photo clusters in their repository.

We have used Bayesian Information Criteria to detect the number of clusters and Jaccard Index to measure the alignment between ground truth and predicted cluster.

We were constrained by the limited amount of real data. The synthetic data fail to give good results because the relationships that real photos have and the subsequent graph patterns are richer than a vague unconnected data. When using real data, clusters are made up of closely related pictures from a personal album with common properties. There are circles or clusters which the model was unable to predict. We have discovered communities that can be explained by the underlying graph structure only. There are other unique and relevant communities which have not been discovered by this model like for instance single photos with no similarity with any existing photo but may form a community with future additions. It would be interesting to explore circle relationships of different users. This method makes it feasible to find communities in personal photos which is very laborious if undertaken manually. The model is completely unsupervised with no user intervention at all. It detects ”Photo Clusters” automatically in personal photo collections. Hierarchical, overlapping and disjoint communities can form and the model takes cognizance of such possibilities. Its a completely unsupervised model and can hence use any kind of personal data from real world other than personal photo collections. Experiments prove the efficacy of the model.

References:

[1]. Maninderjit Kaur and Sushil Kumar Garg, Survey on Clustering Techniques in Data Mining for Software Engineering , International Journal of Advanced and Innovative Research(2278-7844) 3 (2014).
[2]. Xin-JingWang, Lei Zhang, and Ce Liu, Duplicate discovery on 2 billion internet images, IEEE Conference on Computer Vision and Pattern Recognition, CVPR Workshops 2013, Portland, OR, USA, June 23-28, 2013, 2013, pp. 429–436.
[3]. Osama Abu Abbas, Comparisons between data clustering algorithms., Int. Arab J. Inf. Technol. 5 (2008), no. 3, 320–325.
[4]. G Ahalya and Hari Mohan Pandey, Data clustering approaches survey and analysis, Futuristic Trends on Computational Analysis and Knowledge Management (ABLAZE), 2015 International Conference on, IEEE, 2015, pp. 532–537.
[5]. Y.-Y. Ahn, J. P. Bagrow, and S. Lehmann, Link communities reveal multiscale complexity in networks, Nature 466 (2010), 761–764.
[6]. Gergely Palla, Imre Der´enyi, Ill´es Farkas, and Tam´as Vicsek, Uncovering the overlapping community structure of complex networks in nature and society, Nature 435 (2005),no. 7043, 814–818.
[7]. Santo Fortunato and Andrea Lancichinetti, Community detection algorithms: A comparative analysis: Invited presentation, extended abstract, Proceedings of the Fourth International ICST Conference on Performance Evaluation Methodologies and Tools (ICST, Brussels, Belgium, Belgium), VALUETOOLS ’09, ICST (Institute for Computer Sciences Social-Informatics and Telecommunications Engineering), 2009, pp. 27:1–27:2.
[8]. Jaewon Yang and Jure Leskovec, Community-affiliation graph model for overlapping network community detection, Proceedings of the 2012 IEEE 12th International Conference on Data Mining (Washington, DC, USA), ICDM ’12, IEEE Computer Society,2012, pp. 1170–1175.
[9]. Bingyu Liu, Cuirong Wang, Cong Wang, and Yiran Wang, A novel algorithm for finding overlapping communities in networks based on label propagation, Advances in Swarm and Computational Intelligence - 6th International Conference, ICSI 2015, held in conjunction with the Second BRICS Congress, CCI 2015, Beijing, China, June 25-28, 2015,Proceedings, Part II, 2015, pp. 325–332.
[10]. Satu Elisa Schaeffer, Survey: Graph clustering, Comput. Sci. Rev. 1 (2007), no. 1, 27–64.
[11]. E. Ravasz and A. L. Barab´asi, Hierarchical organization in complex networks, Physical Review E 67 (2003), 026112.
[12]. Jonathan Chang and David M. Blei, Relational topic models for document networks.,
[13]. AISTATS (David A. Van Dyk and Max Welling, eds.), JMLR Proceedings, vol. 5, JMLR.org, 2009, pp. 81–88.
[14]. Aditya Krishna Menon and Charles Elkan, Link prediction via matrix factorization, Proceedings of the 2011 European Conference on Machine Learning and Knowledge Discovery in Databases - Volume Part II (Berlin, Heidelberg), ECML PKDD’11, Springer-Verlag, 2011, pp. 437–452.
[15]. Duy Quang Vu, Arthur U. Asuncion, David R. Hunter, and Padhraic Smyth, Dynamic egocentric models for citation networks., ICML (Lise Getoor and Tobias Scheffer, eds.),Omnipress, 2011, pp. 857–864.
[16]. Tetsuya Yoshida, Toward finding hidden communities based on user profile, ICDMW 2010, The 10th IEEE International Conference on Data Mining Workshops, Sydney, Australia, 13 December 2010, 2010, pp. 380–387.
[17]. Julian Mcauley and Jure Leskovec, Discovering social circles in ego networks, ACM Trans. Knowl. Discov. Data 8 (2014), no. 1, 4:1–4:28.
[18]. Stephen C. Johnson, Hierarchical clustering schemes, Psychometrika 32 (1967), no.3, 241–254.
[19]. Edoardo M. Airoldi, David M. Blei, Stephen E. Fienberg, and Eric P. Xing, Mixed membership stochastic blockmodels, J. Mach. Learn. Res. 9 (2008), 1981–2014.

[20]. Ramnath Balasubramanyan and William W. Cohen, Block-lda: Jointly modeling entityannotated text and entity-entity links., SDM, SIAM / Omnipress, 2011, pp. 450–461.
[21]. Bhalla V., Chaudhury S. (2017) Artificial Immune Hybrid Photo Album Classifier. In: Raman B., Kumar S., Roy P., Sen D. (eds) Proceedings of International Conference on Computer Vision and Image Processing. Advances in Intelligent Systems and Computing, vol 459. Springer, Singapore. https://doi.org/10.1007/978-981-10-2104-6_43