Information Enhancement for Travelogues via a Hybrid Clustering Model

Lu Zhang, Jingsong Xu, Jian Zhang, Yongshun Gong
Global Big Data Technologies Centre, University of Technology Sydney, Sydney, Australia
Email: {Lu.Zhang-5@student., Jingsong.Xu@, Jian.Zhang@, Yongshun.Gong@student.} uts.edu.au

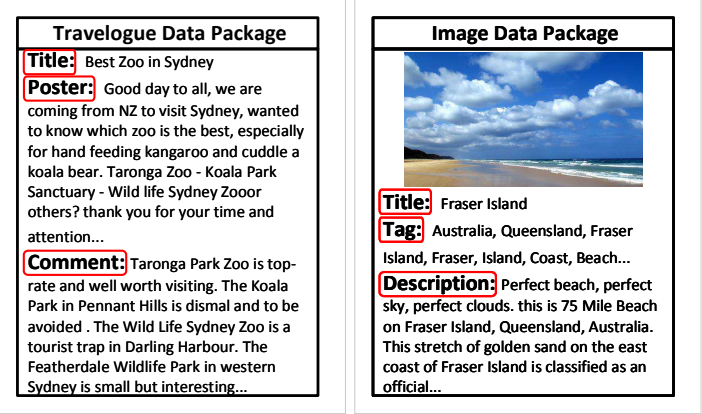
***Abstract*—Travelogues consist of textual information shared by tourists through web forums or other social media which often lack illustrations (images). In image sharing websites like Flicker, users can post images with rich textual information: 'title', 'tag' and 'description'. The topics of travelogues usually revolve around beautiful sceneries. Corresponding landscape images recommended to these travelogues can enhance the vividness of reading. However, it is difficult to fuse such information because the text attached to each image has diverse meanings/views. In this paper, we propose an unsupervised Hybrid Multiple Kernel K-means (HMKKM) model to link images and travelogues through multiple views. Multi-view matrices are built to reveal the correlations between several respects. For further improving the performance, we add a regularisation based on textual similarity. To evaluate the effectiveness of the proposed method, a dataset is constructed from TripAdvisor and Flicker to find the related images for each travelogue. Experiment results demonstrate the superiority of the proposed model by comparison with other baselines.**

***Index Terms*—multiple kernel k-means, multi-view clustering, multimedia analyses, information enhancement**

I. INTRODUCTION

Due to the rapid development of the Internet, large volumes of multi-media data enrich travel experiences of people [1]. A large number of trip sharing websites provide sufficient travelogues for tourists who are planning their future trips. Most of such trip sharing websites (e.g. TripAdvisor and Wikitravel) only provide heaps of humdrum textual contents without vividness and abundance. On the other hand, by some visual image sharing websites (e.g. Flickr and Pinterest) it is much more convenient for people to get plenty of vivid images. But people cannot gain precise and detailed information by mere images. Therefore, the cross-domain information can be connected by the same topic. These data have different emphases and formats, which can enhance the reading experience. With the illustrated travelogues, people could feel the majestic sceneries directly rather than be confused about the tedious words.

In order to integrate vivid illustration into textual travelogues, we need to bridge gaps between different domain information. In particular, there are three major difficulties. (1) **Heterogeneous features**. Features are represented heterogeneously regarding their modalities, which brings difficulties for mining inner correlations between images and texts. (2) **Restricted topics**. Considering the specificity of travel sharing data in some specific countries or places, topics are more restricted compared with public datasets. While in some other public datasets with images and texts, such as NUS-WIDE [2], Pascal VOC 2007 [3], ImageNet [4], Wiki [5] and MIT Place 205 [6], topics can be of very wide range. It could be easier to recognise topics between multiple fields, such as sports news and political news than to distinguish relatively fine distinctions in one field. (3) **Insufficiency of textual information**. Texts in these public datasets are often quite short. A few tags or simple sentences related to the contents of images can only be seen as the supplementary information in the same domain.



(a) (b)

Fig. 1: Travelogue data examples. (a) Each travelogue is composed by three parts: 'title', 'poster' and 'comment'. (b) Image data example. Each image has three kinds of textual information: 'title', 'tag' and 'description'.

We build a new multi-view dataset which contains both textual travelogues and landscape images (crawled from TripAdvisor and Flickr), as shown in Fig. 1. We manually label travelogues with matched images as our ground truth. Specifically, each image includes three views of explanatory textual information termed 'title', 'tag' and 'description'. Every view has its emphasis as complementary information of other views, which expands textual description of images.

The objective is to link information (images and texts) between two domains. One traditional solution is to learn transformation matrices to maximise the similarity, like Canonical Correlation Analysis (CCA) [7]. However, there are already a lot of semantic gaps between the image content and textual information attached to the images in Flicker. The problem

is that it would be much more difficult to match textual travelogues and images by learning the map between them directly. In turn, considering images from Flicker have multi-view textual information (title, tag, and description), in this paper, we propose to match travelogues and images through textual information directly. Since title, tag and description can represent the image from different views, Multiple Kernel K-means (MKKM) [8] based model is adopted to discover Top-N correlative images for travelogues. We apply different word embedding methods, such as Term Frequency-Inverse Document Frequency [9]–[11] and Word2Vec [12], [13] to build kernel matrices and propose a hybrid MKKM model. At first, a multi-view similarity framework is built to reveal the correlations from several perspectives/views. The HMKKM method is used to mine potential associations among these views. To improve the performance of this unsupervised learning process, a regularisation is also introduced to construct a hybrid model. We conduct experiments to evaluate the proposed method and compare with other baselines on a dataset constructed from TripAdvisor and Flickr. The results show the superiority of the proposed method against other compared models.

The rest of the paper is organised as follows. In Section 2, we review some related work. Section 3 introduces our cross-domain hybrid model. Section 4 focuses on the experiments settings and results. In section 5, we make a conclusion.

II. RELATED WORK

Several text representation methods have been proposed in previous work [10]–[14]. Term Frequency-Inverse Document Frequency (TF-IDF) [9]–[11], [14] is one of the most classic algorithms. TF-IDF directly represents one article by TF-IDF scores. It grades each term by considering both occurrence times and importance. A x-dimensional vector can be built for each article, if totally x terms are meaningful in the whole corpus. It has been adopted as a baseline in many papers [15], [16] due to its simplicity.

Another word embedding model Word2Vec [12], [13] is widely adopted in many works since it was proposed [17], [18]. Reference [17] used Word2Vec to embed texts as vectors for verifying the integrity of images and associated texts in social network. As a variant of Word2Vec, Doc2Vec [13] embeds long paragraphs in latent space. Latent Dirichlet Allocation (LDA) [19] is another popular algorithm for word embedding. LDA is widely used in cross-modal cross-domain multimedia information retrieval and mining [5], [20]–[23].

Aforementioned models map text information into vector space which makes good preparation for further semantic analysis. But TF-IDF can just catch the same words in different documents. It does not put synonyms into consideration [24]. Word2Vec considers context information by introducing CBOW and N-gram model. LDA tries to extract topics of articles by supposing a Dirichlet allocation of topics. Word2Vec and LDA aim to capture in-depth semantic information but can not deal with multi-view features of one document.

Canonical Correlation Analysis (CCA) [7], [25] is a representative statistic method for exploring the relationship between two sets of variables. It aims at maximizing the correlation of two related multi-domain data. Similar to CCA, Partial Least Squares (PLS) [26] is another classic method which aims to learn a linear projection that maps different domains into a common latent subspace. Reference [26] utilised PLS for multi-modal face recognition. In reality, the linear projection is not applicative in many cases which may lead to limited performance [27].

Some deep learning based methods [28], [29] have shown their excellent performance. Reference [27] proposed a cross-modal correlation learning approach with multi-grained fusion by a hierarchical network. It divided the task into two steps and used deep belief network for coarse-grained learning. Reference [16] proposed an approach for text illustration from a tagged repository. Reference [30] proposed a generalized semi-supervised structured subspace learning model for cross-modal retrieval. However, these supervised or semi-supervised learning methods need paired data for training which are not applicable to our small dataset.

Different from above-mentioned methods, in this paper, we take the advantage of image properties in Flicker, where each image has abundant textual information: title, tag, and description. Each property can represent the image in different view. The problem of mining relations between textual travelogues and landscape images can be seen as mining relations between textual information through multiple views. We build a hybrid multiple kernel k-means model using different textual features for clustering travelogues and images without a large number of training samples. Our proposed model can reveal the direct relations among texts by fusing multi-view features. We also show that the performance of the proposed method can be further improved by introducing a regularisation based on overall text similarity.

III. HYBRID MULTIPLE KERNEL K-MEANS MODEL

Detailed textual travelogues illustrated by vivid scenery images can provide rich and varied information, thus helping readers make decisions effectively and rapidly. The interactions of the multi-view content can not only enhance the enrichment but also ease the work of authors who may spend time on collecting pictures according to their destinations in one text by automatically recommending the most relevant vivid images from a slot.

For these reasons, we build a dataset which contains textual travelogues and images as two parts (see Fig. 1). Each image has three kinds of textual information as three views of complementary textual information: title, tag, and description. Every view has its emphasis, which expands and enhances the textual description of images. Three views of image textual information give us clues on using a multi-view learning method. Different views contain different emphases which complement each other. Single view can not leverage the abundant information which promotes accuracy.

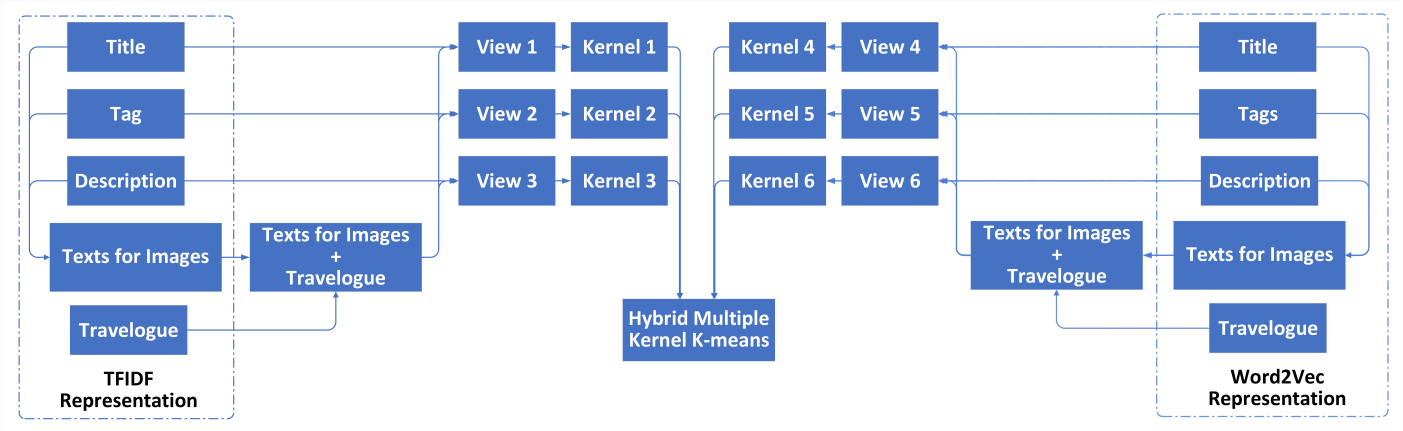


Fig. 2: Framework of the hybrid multiple kernel k-means (HMKKM) model. Two basic components compose the model. The left one is based on TF-IDF representation method and the right one is based on Word2Vec representation method.

There are mainly four steps in the proposed method. Firstly, all travelogues and images are represented by TF-IDF and Word2Vec encoding methods respectively. Based on the word encoding, multi-view matrices are constructed by title, tag, and description of images. In this process, each travelogue and image is re-represented again. Then multi-view matrices are adopted to construct kernel similarity matrices. Finally, the proposed HMKKM is used to cluster travelogues and images for final recommendation.

A. Kernel matrix construction

The proposed hybrid multiple kernel k-means (HMKKM) model aims to match travelogues and images together in four steps by mining correlations between them. Suppose we collect dataset $\mathcal{D} = \{d_1, \cdots, d_j, d_{j+1}, \cdots, d_{j+r}\}$ which contains all textual components of j images and r travelogues. That is, for each image, title, tag, and description are collected together, as well as title, poster and comment for each travelogue. Given a travelogue d_q $(j < q < j + r + 1)$, the goal of HMKKM is to find the Top-N closest landscape images d_u $(0 < u < j + 1)$.

Firstly, TF-IDF and Word2Vec are adopted to separately encode all texts in $\mathcal{D}$ as $W = [\mathbf{w}_1^\top; \cdots; \mathbf{w}_j^\top; \mathbf{w}_{j+1}^\top; \cdots; \mathbf{w}_{j+r}^\top]$. Based on this, in the second step, each row vector in W is further represented as three vectors in three feature space $V^s = [\mathbf{v}_1^{s\top}; \cdots; \mathbf{v}_j^{s\top}; \mathbf{v}_{j+1}^{s\top}; \cdots; \mathbf{v}_{j+r}^{s\top}] \in \mathbb{R}^{(j+r)\times j}, \forall s \in \{1, 2, 3\}$. As mentioned before, each image consists of three views of textual information which can be encoded as $\mathbf{w}_{t_i}$, $\mathbf{w}_{g_i}$, $\mathbf{w}_{e_i}$ by TF-IDF or Word2Vec separately. t_i, g_i, e_i represent the title, tag and description respectively of image i. Totally there are j titles/tags/descriptions. Each row vector $\mathbf{w}_h^\top$ $(0 < h < j + r + 1)$ in W is further embedded as three different vectors $\mathbf{v}_h^{1\top}$, $\mathbf{v}_h^{2\top}$, $\mathbf{v}_h^{3\top}$ which are constructed by title, tag and description separately. Suppose by tag, o-th element $\mathbf{v}_{h,o}^s$ $(0 < o < j+1)$ is cosine similarity of document h and tag of o-th image:

$$\mathbf{v}_{h,o}^s = \frac{\mathbf{w}_h^\top * \mathbf{w}_{g_o}}{\|\mathbf{w}_h\|_2 \|\mathbf{w}_{g_o}\|_2} \tag{1}$$

Suppose TF-IDF is chosen for encoding, three view matrices of size $(j + r) \times j$ are built in which each row represents each document, each column represents title/tag/description of each image. Namely title, tag, and description are regarded as three bases to construct the view matrices. Concerning two encoding methods, in total six view matrices are built.

Thirdly, we construct kernel matrices by row vectors in V^s. Based on the six view matrices of size $(j + r) \times (j + r)$, it is easy to build six kernel matrices with predefined kernel functions such as Gaussian kernel or linear kernel. Different views contain different emphases which complement each other.

HMKKM model consists of two same components as shown in Fig. 2. The left block is TF-IDF based kernel matrices generation component using TF-IDF to represent text information. The right block is Word2Vec based kernel matrices generation component. The diverging kernel matrices generated by both basic components are fed into the proposed HMKKM algorithm.

B. Hybrid multiple kernel k-means method

There are totally $j + r$ samples in our dataset. Assume we only use single kernel, $\phi(\cdot)$ is the mapping function that projects samples into the new kernel Hilbert space. The sum-of-squares loss of kernel k-means algorithm is [8]:

$$\begin{aligned} \min_{Z\in\{0,1\}^{(j+r)\times k}} & \sum_{i=1,c=1}^{j+r,k} Z_{ic}\|\phi(\mathbf{v}_i) - \mu_{\mathbf{c}}\|_2^2 \\ \text{s.t.} & \sum_{c=1}^{k} Z_{ic} = 1 \end{aligned} \tag{2}$$

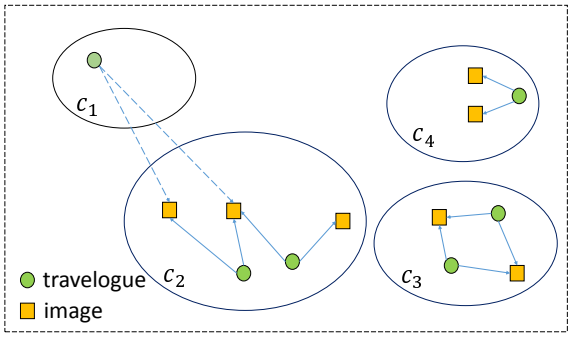


Fig. 3: Recommendation strategy in HMKKM model. Travelogues and images in the same circle belong to the same class. For travelogues in class c_2, recommend Top-N nearest images for these travelogues within the same class. If there is less than N images in the same class like the travelogue in class c_1, images in the nearest class c_2 will be chosen.

In Eq.(2), $\mathbf{Z} \in \{0,1\}^{(j+r)\times k}$ is the clustering assignment matrix. If sample i belongs to cluster c, $Z_{ic} = 1$, otherwise $Z_{ic} = 0$. $\mu_\mathbf{c}$ is the centroid of the c-th cluster:

$$\mu_c = \frac{1}{n_c}\sum_{i=1}^{j+r} Z_{ic}\phi(\mathbf{v}_i) \tag{3}$$

n_c is the total samples number in the c-th cluster:

$$n_c = \sum_{i=1}^{j+r} Z_{ic} \tag{4}$$

If $\mathbf{K}$ is a kernel matrix with:

$$K_{ij} = \phi(\mathbf{v}_i)^\top \phi(\mathbf{v}_j) \tag{5}$$

and:

$$\mathbf{G} = \mathsf{diag}([n_1^{-1}, n_2^{-1}, \cdots, n_k^{-1}]) \tag{6}$$

By defining $\mathbf{H} = \mathbf{Z}\mathbf{G}^{\frac{1}{2}}$, Equation (2) can be relaxed as the following format:

$$\begin{aligned} &\min_{\mathbf{H}} \mathrm{Tr}(\mathbf{K}(\mathbf{I}_{j+r} - \mathbf{H}\mathbf{H}^\top)) \\ &\text{s.t. } \mathbf{H} \in \mathbb{R}^{(j+r)\times k}, \mathbf{H}^\top\mathbf{H} = \mathbf{I}_k \end{aligned} \tag{7}$$

where $\mathbf{I}_k$ is a $k \times k$ identity matrix.

In multiple kernel k-means clustering (MKKM), each sample can be mapped as $\phi_{\boldsymbol{\beta}}(\mathbf{v}) = [\beta_1\phi_1(\mathbf{v})^\top, \cdots, \beta_m\phi_m(\mathbf{v})^\top]^\top$, in which $\boldsymbol{\beta} = [\beta_1, \cdots, \beta_m]^\top$ is the coefficients vector. Then (7) can be rewritten as:

$$\begin{aligned} &\min_{\mathbf{H},\boldsymbol{\beta}} \mathrm{Tr}(\mathbf{K}_{\boldsymbol{\beta}}(\mathbf{I}_{j+r} - \mathbf{H}\mathbf{H}^\top)) \\ &\text{s.t. } \mathbf{H} \in \mathbb{R}^{(j+r)\times k}, \mathbf{H}^\top\mathbf{H} = \mathbf{I}_k,\ \boldsymbol{\beta}^\top\mathbf{1}_m = 1, \beta_p \geq 0, \forall p. \end{aligned} \tag{8}$$

$\mathbf{1}_m \in \mathbb{R}^m$ is a column vector with all 1 elements.

In the proposed model, we formulated the model in this form:

$$\begin{aligned} &\min_{\mathbf{H},\boldsymbol{\beta}} \mathrm{Tr}(\mathbf{K}_{\boldsymbol{\beta}}(\mathbf{I}_{j+r} - \mathbf{H}\mathbf{H}^\top)) + \alpha\mathrm{Tr}(\mathbf{H}^\top\mathbf{L}\mathbf{H}) \\ &\text{s.t. } \mathbf{H} \in \mathbb{R}^{(j+r)\times k}, \mathbf{H}^\top\mathbf{H} = \mathbf{I}_k,\ \boldsymbol{\beta}^\top\mathbf{1}_m = 1, \beta_p \geq 0, \forall p. \end{aligned} \tag{9}$$

TABLE I: The best results with different numbers of kernels on validation set.

Number of kernels	MAP@10
2	0.3818
3	**0.4016**
4	0.3870
5	0.3710
6	0.3643

The first term is the traditional multiple kernel k-means objective function. m is the number of views (kernels). Followed by an regularisation term in which we construct a graph Laplacian matrix $\mathbf{L}$, defined as: $\mathbf{L} = \mathbf{D} - \mathbf{W}$. $\mathbf{W}$ is a graph adjacency matrix constructed from the similarity result from TF-IDF, and $\mathbf{D}$ is the diagonal degree matrix $\mathbf{D}_{ii} = \sum_j(\mathbf{W}_{ij})$. The objective of this regularisation function is to utilise the similarities between textural travelogues and images generated by TF-IDF to guide the clustering process. Parameter α controls the trade off between model complexity and empirical loss. Since kernels are adopted in our method, a number of kernels, e.g. linear kernel, Gaussian kernel can be directly applied.

This problem can be solved by alternately updating $\mathbf{H}$ and $\boldsymbol{\beta}$ [8]. Firstly optimise $\mathbf{H}$ with fixed $\boldsymbol{\beta}$. With the kernel coefficients $\boldsymbol{\beta}$ fixed, $\mathbf{H}$ can be obtained by choosing the l smallest eigenvectors of $(-\mathbf{K}_{\boldsymbol{\beta}} + \alpha\mathbf{L})$. Secondly, optimise $\boldsymbol{\beta}$ with fixed $\mathbf{H}$. With $\mathbf{H}$ fixed, $\boldsymbol{\beta}$ can be optimised via solving a quadratic programming probelm with linear constraints:

$$\beta_p = \frac{\mathrm{Tr}(K_p(\mathbf{I} - \mathbf{H}\mathbf{H}^\top))^{-1}}{\sum_{p=1}^{m}(\mathrm{Tr}(K_p(\mathbf{I} - \mathbf{H}\mathbf{H}^\top))^{-1})} \tag{10}$$

After clustering by HMKKM, all travelogues and images are clustered as shown in Fig. 3. Suppose that the target travelogue is labeled as class c_2, the Top-N nearest images belonging to this class will be recommended to the target travelogue. If a target travelogue is in class c_1 which contains insufficient images, a greedy strategy is used to find the nearest class c_2 for recommendation until Top-N requirement is satisfied (as dashed line shown in the Fig. 3).

IV. EXPERIMENT RESULTS

A. Experiment Description

Dataset. To evaluate our approach, a new dataset is constructed. We search on the TripAdvisor website which is one of the most popular trip information sharing websites in the world and provides interactive travel forums. Travelogues that contain too many mistakes and advertisements will be discarded. The length of each travelogue is also controlled by constraining the reply number of each poster is more than 50. In this way, we collect 125 travelogues in Australia Travel

Forum as 125 articles from 2016 to 2017. Each travelogue is composed of three parts: title, poster, and comment as shown in Fig. 1a.

We also use the keywords in the travelogue to search images on Flickr. For the information effectiveness, images with less than 10 words description will be filtered, resulting totally 100 images. Each image has three kinds of textual information as three views of complementary textual information: title, tag, and description (see Fig. 1b).

Evaluation. The performance is measured by MAP@N score [5], [18], [21] where we set $N = 1, 5, 10, 15, 20$. Mean Average Precision (MAP) is a widely used evaluation metric in recommendation system and information retrieval. For a recommendation system, precision (P) is defined as:

$$P = \frac{TP}{TP + FP}$$

where TP refers to the number of our recommendations that are relevant, FP refers to the number of our recommendations that are not relevant. Precision at cutoff k ($P(k)$) is P only considering the subset of recommendations from rank 1 to k. Definition of Average Precision (AP) is:

$$AP@N = \frac{1}{\min(m, N)} \sum_{k=1}^{N} P(k)$$

Here m is the number of relevant items. N is the number of recommendation items. If $k-th$ item is not relevant, $P(k) = 0$. MAP@N is given by:

$$MAP@N = \frac{\sum_{q=1}^{Q} AP@N}{Q}$$

where Q is the number of query.

Baselines. Five baselines are used to compare with our proposed model: TF-IDF [9]–[11], Word2Vec [12], [13], Doc2Vec [13], PLS [26], and CCA [7]. All the textual information of each image will be combined and represented by TF-IDF, Word2Vec, and Doc2Vec respectively as well as each travelogue. For TF-IDF, we collect top 1004 important words with high term frequency as a dictionary. Each text is mapped as a 1004-dimension vector. For Word2Vec, we use a pre-trained model with Google News corpus [31] to embed each text into a 300-dimension vector. For Doc2Vec, the model is pre-trained on Lee Background Corpus [32] which embeds each text as a 50-dimension vector. For these three methods, without clustering step, we just compare the similarities between travelogues and images using cosine distance directly. For PLS and CCA, all view matrices are mapped linearly as six new matrices by making multiple views of the same sample closer. We then average the mapped matrices to get one matrix whose rows are embedded vectors of documents in dataset $\mathcal{D}$. Then we calculate cosine similarity.

B. Parameter analysis and experiment results

Parameter analysis. We randomly choose 25 documents from travelogues in dataset $\mathcal{D}$ and all the image textual information as validation set. The clustering number k and the regularisation parameter α are set to 22 and 10 respectively to achieve the best performance after a line search process. Since there are total 6 views adopted in the model we firstly evaluate and select the best combination of kernels. Gaussian kernel is adopted to project the original space into high-dimensional latent space. The experimental results are shown in Table I. It is clear to see that using 3 kernels can obtain the best performance. Specifically, TF-IDF is selected for title and description views, while Word2vec is for tag view. The reason is that TF-IDF only catches the same words. Additionally, Word2vec takes synonyms into consideration. A title always shows topic of one text which is the most important information. TF-IDF captures this important word directly. In consider of tag view, tag information is not more important than title information and contains fewer words than description. By TF-IDF, the result matrix will be too sparse to be effective. For the description view, if we adopt Word2Vec, the result will be very diverging because Word2Vec considers synonyms as well. We finally use this option and compare with other methods.

Experiment results and discussion. Evaluation results are shown in Table II. Besides five baselines, firstly we use single encoding method to generate three kernel matrices with Gaussian kernel (TFIDF-MKKM and Word2Vec-MKKM) and feed these into our HMKKM model. Secondly, we use both two encoding schemes to generate six kernel matrices with Gaussian kernel and linear kernel. From the results, it is easy to observe that our proposed HMKKM method achieves significant results on the dataset. In detail, our HMKKM with Gaussian kernel achieves 61.60%, 47.42%, 40.27%, 35.41% and 34.02% at MAP@1, MAP@5, MAP@10, MAP@15, and MAP@20. It outperforms all the baselines with a large margin. HMKKM with linear kernel is the second best which also outperforms all the baselines. This shows that our proposed model achieves top and stable performances.

In Doc2Vec, Word2Vec and TF-IDF models, only single encoding method is used which leads to inferior performance. Word2Vec and TF-IDF models perform better than Doc2Vec. This shows that, for relatively short text representation, Word2Vec and TF-IDF are more effective which can also explain why we choose these as our encoding methods in our step one. The performance of Doc2Vec is inferior to other methods. MAP@10, MAP@15, and MAP@20 are the worst even below 10%. The reason is that both travelogues and images have short texts, while Doc2Vec is more effective in long text mapping. Meanwhile, the encoded feature vector trained from Lee Background Corpus [32] is not suited to be applied in this typical area. This indicates that only single word representation method can not mine real correlations between our data and may miss important information that is needed for matching.

PLS and CCA achieve similar performance, whose MAP scores are above 11% and below 29%. It is reasonable that these two methods are proposed for cross-modal information correlation mining. In our case, both texts from Flickr and TripAdvisor are involved which contains minor gaps compared

title: Best Zoo in Sydney
poster: Good day to all, we are coming from NZ to visit Sydney, wanted to know which zoo is the best, especially for hand feeding kangaroo and cuddle a koala bear.- Taronga Zoo- Koala Park Sanctuary- Wild life Sydney Zooor others? thank you for your time and attention...
comment: Taronga Park Zoo is top-rate and well worth visiting. The Koala Park in Pennant Hills is dismal and to be avoided. The Wild Life Sydney Zoo is a tourist trap in Darling Harbour. The Featherdale Wildlife Park in western Sydney is small but interesting...

title: barn swallow

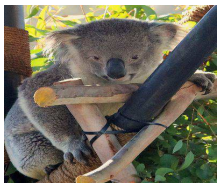
tag: yoho national park, emerald lake, barn, swallow, birds, bird, wild, wildlife, outdoors

description: at the borders of emerald lake while enjoying...

title: heading towards a safe

tag: happy holidays, predators, mammals, zoogdieren, leeuwin, cub, young, lioness, pantheraleo, carnivore, natuur, nature, savannah, grass, masaimara, eastafrica, oostafrika, kenya, backlight, tegenlicht, merrychristmas

description: dsc taken in wild kenya i would like to thank all of you for your visits to my photostream and for your comments and invites they are...



title: koala well hello there

tag: cute animals, zoo, wildlife, zoo animals, zoo photography, zoo images, koala, animal world

description: the koala is an arboreal herbivorous mammal and native to australia...

title: koala park

tag:

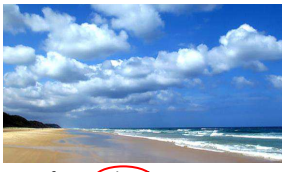
description: west pennant hills a suburb in the hills district of sydney new south wales australia it is sad to hear bad news about the park when we visited in it was well kept we did not see any neglect of animals the koala park...

title: pinnacles bobcat death stare

tag: national, national park, pinnacles, rocks, cave, bobcat, wild, animal, cat, kitten, kitty,

description: we were in livermore for a family christmas event and had to take the looooooong drive home to the ie we did the whole ever so exciting...

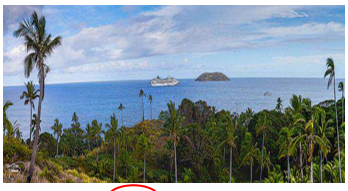
title: Pacific Coast Drive
poster: I am interested in visiting Palm Beach while doing this drive. Is is en route or would it make more sense to take a day trip from Sydney? Thank you.
comment: Pacific Coast Drive - see link – is the Pacific Highway between Sydney and Brisbane – distance is 1,000km (it's M1 and A1 on google maps). Palm Beach is not on the Pacific Highway – it is a northern beachside suburb of Sydney and is 45km from the city – look at google maps – just a do a day trip from the city.



title: fraser island

tag: australia, queensland, fraserisland, island, coast, beach

description: explored perfect beach perfect sky perfect clouds this is mile beach on fraser...



title: dravuni island

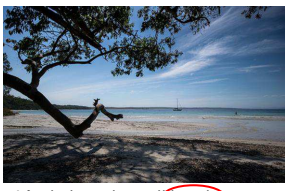
tag: beach, beautiful, fiji, island, sea, ship, south pacific

description: this island is void of cars stores in fact nothing to say that were living in the century i climbed a mountain to take this shot many visitors come to snorkel in the great astrolabe reef or to hike dravuni island's highest peak for magnificent...

title: new zealand christmas weather

tag:

description: it was obviously beach weather our son and his fiancee and our youngest daughter and her husband were already at pukerua bay a small beach we found as we drove down the steep narrow road towards the beach...



title: hole in the wall beach jervis bay

tag: jervisbay, hole in the wall, park, shark, sailing, beach, ocean, sea, fishing, sail, australia, summer, colour, holiday, new south wales, nsw, australia, white, sand

description: located in the booderee national park at jervis bay hole in the wall beach is just lovely hyams beach on this day both of those beaches...

title: bronzed aussie

tag: melbourne, beach, sun, summer, sand, bathers, red, blue, towel, man, boy, sunbake, sunbathing, towels, sunscreen, aussie

description: how the locals cope with a new years eve no raising of the aussie flag though

Fig. 4: Two examples of travelogue enhancement. For each one, travelogue information is in the left top. Five images around this travelogue are the five results of Top-10 recommendation by HMKKM model. Keywords are highlighted in red ellipses.

TABLE II: Travelogue visualization performance.

Method	MAP@1	MAP@5	MAP@10	MAP@15	MAP@20
Doc2Vec	0.2320	0.1398	0.0979	0.0894	0.0918
Word2Vec	0.4400	0.3402	0.3111	0.2932	0.2842
TF-IDF	0.4800	0.3933	0.3427	0.3169	0.3116
PLS	0.2160	0.1676	0.1442	0.1310	0.1279
CCA	0.2880	0.1783	0.1360	0.1232	0.1215
TFIDF-MKKM	**0.6160**	0.4499	0.3489	0.2929	0.2843
Word2Vec-MKKM	0.5520	0.4407	0.3581	0.3476	0.2732
HMKKM (with linear kernel)	0.5680	0.4674	0.3896	0.3469	0.3320
HMKKM (with Gaussian kernel)	**0.6160**	**0.4742**	**0.4027**	**0.3541**	**0.3402**

with cross-modal data.

TFIDF-MKKM and Word2Vec-MKKM achieve similar performance too. Results of TFIDF-MKKM and Word2Vec-MKKM are better than other baselines only except MAP@15 and MAP@20 of Word2Vec and TF-IDF. This proves the effectiveness of our framework. Additionally, the results are not better than HMKKM with Gaussian and linear kernel which shows the superior performance of our proposed hybrid model. It also shows that Gaussian kernel can only obtain slightly higher performance, making linear kernel suitable for practical usage.

Fig. 4 shows two examples of travelogue illustration results from HMKKM model. For each one, travelogue information is at the left top which contains three parts as text: title, poster, and comment. Five images around the travelogue are the five results of Top-10 recommendation by HMKKM model. These five images all include three views of textual information: title, tag, and description. As we can easily find, in the first example, this travelogue is about 'zoo', 'Sydney', 'koala', 'wildlife', 'park' and 'Pennant Hills'. From these keywords in the red ellipses of image textual information we can see, same words appear, such as 'zoo', 'koala', 'Sydney', 'wild life' and 'park'. In the second example, same words such as 'beach', 'coast' can also be found in both travelogue and images. TF-IDF method is good at catching these completely same words. Additionally, word pairs with similar meaning but not the same word such as 'animal'-'wild life', 'Australia'-'Sydney', 'nature'-'wild' in the first example and 'drive'-'car', 'highway'-'road' in the second example can be mapped closely in target space by Word2Vec. Both two representation methods have their different emphases which shows good evidence of the effectiveness of our proposed HMKKM model by merging TF-IDF generalized kernel and Word2Vec generalized kernel.

V. CONCLUSION

In this paper, we propose a novel unsupervised multi-view hybrid multiple kernel k-means (HMKKM) model to tackle travelogue visualization problem. We collect a cross-domain media dataset which can be used for further research on multimedia data matching. We use the structure information which complements each other to build multiple view matrices. For further improving the performance, a regularization based on textual similarity is introduced in HMKKM. The results of experiments show that the proposed HMKKM model can mine correlation of travelogues and landscape images effectively. Monotonous textual information is enhanced by vivid images through our proposed model. In the future, we will evaluate the model on more multimedia datasets and explore more meaningful applications.

REFERENCES

[1] X. Lu, Y. Pang, Q. Hao, and L. Zhang, "Visualizing textual travelogue with location-relevant images," in *Proceedings of the international workshop on Location Based Social Networks*, 2009, pp. 65–68.

[2] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y. Zheng, "Nus-wide: a real-world web image database from national university of singapore," in *Proceedings of the ACM international conference on image and video retrieval*. ACM, 2009, p. 48.

[3] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes challenge 2007 (voc 2007) results (2007)," http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html, 2008.

[4] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the devil in the details: Delving deep into convolutional nets," *arXiv preprint arXiv:1405.3531*, 2014.

[5] N. Rasiwasia, J. Costa Pereira, E. Coviello, G. Doyle, G. R. Lanckriet, R. Levy, and N. Vasconcelos, "A new approach to cross-modal multimedia retrieval," in *Proceedings of the ACM international conference on Multimedia*, 2010, pp. 251–260.

[6] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, "Learning deep features for scene recognition using places database," in *Advances in Neural Information Processing Systems*, 2014, pp. 487–495.

[7] H. Hotelling, "Relations between two sets of variates," *Biometrika*, vol. 28, no. 3/4, pp. 321–377, 1936.

[8] X. Liu, M. Li, L. Wang, Y. Dou, J. Yin, and E. Zhu, "Multiple kernel k-means with incomplete kernels," in *Proceedings of the Association for the Advancement of Artificial Intelligence*, 2017, pp. 2259–2265.

[9] G. Salton, E. A. Fox, and H. Wu, "Extended boolean information retrieval," *Communications of the ACM*, vol. 26, no. 11, pp. 1022–1036, 1983.

[10] K. Sparck Jones, "A statistical interpretation of term specificity and its application in retrieval," *Journal of documentation*, vol. 28, no. 1, pp. 11–21, 1972.

[11] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Information processing & management*, vol. 24, no. 5, pp. 513–523, 1988.

[12] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[13] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in Neural Information Processing Systems*, 2013, pp. 3111–3119.

[14] H. Luhn, "A statistical approach to mechanized encoding," *IBM journal of Research and Development*, 1957.

[15] L. Pang, Y. Lan, J. Guo, J. Xu, S. Wan, and X. Cheng, "Text matching as image recognition," in *Proceedings of the Association for the Advancement of Artificial Intelligence*, 2016, pp. 2793–2799.

[16] H. Jhamtani, S. Varma, M. Gundapuneni, and S. K. Dutta, "A supervised approach for text illustration," in *Proceedings of the ACM international conference on Multimedia*, 2016, pp. 217–221.

[17] A. Jaiswal, E. Sabir, W. AbdAlmageed, and P. Natarajan, "Multimedia semantic integrity assessment using joint embedding of images and text," in *Proceedings of the ACM international conference on Multimedia*, 2017, pp. 1465–1471.

[18] Z. Ren, H. Jin, Z. Lin, C. Fang, and A. Yuille, "Joint image-text representation by gaussian visual-semantic embedding," in *Proceedings of the ACM international conference on Multimedia*, 2016, pp. 207–211.

[19] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, no. Jan, pp. 993–1022, 2003.

[20] S. Qian, T. Zhang, and C. Xu, "Multi-modal multi-view topic-opinion mining for social event analysis," in *Proceedings of the ACM international conference on Multimedia*, 2016, pp. 2–11.

[21] S. Qian, T. Zhang, R. Hong, and C. Xu, "Cross-domain collaborative learning in social multimedia," in *Proceedings of the ACM international conference on Multimedia*, 2015, pp. 99–108.

[22] C. Kang, S. Xiang, S. Liao, C. Xu, and C. Pan, "Learning consistent feature representation for cross-modal multimedia retrieval," *IEEE Transactions on Multimedia*, vol. 17, no. 3, pp. 370–381, 2015.

[23] M. Fan, W. Wang, P. Dong, L. Han, R. Wang, and G. Li, "Cross-media retrieval by learning rich semantic embeddings of multimedia," in *Proceedings of the ACM international conference on Multimedia*, 2017, pp. 1698–1706.

[24] F. S. W. Y. P. H. Yazhou Yao, Jian Zhang and Z. Tang, "Discovering and distinguishing multiple visual senses for polysemous words," in *Proceedings of the Association for the Advancement of Artificial Intelligence*, pp. 523–530.

[25] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor, "Canonical correlation analysis: An overview with application to learning methods," *Neural computation*, vol. 16, no. 12, pp. 2639–2664, 2004.

[26] A. Sharma and D. W. Jacobs, "Bypassing synthesis: Pls for face recognition with pose, low-resolution and sketch," in *Proceedings of the conference on Computer Vision and Pattern Recognition*, 2011.

[27] Y. Peng, J. Qi, X. Huang, and Y. Yuan, "Ccl: Cross-modal correlation learning with multigrained fusion by hierarchical network," *IEEE Transactions on Multimedia*, vol. 20, no. 2, pp. 405–420, 2018.

[28] H. Fan, X. Chang, D. Cheng, Y. Yang, D. Xu, and A. G. Hauptmann, "Complex event detection by identifying reliable shots from untrimmed videos," in *Proceedings of the IEEE International Conference on Computer Vision*. IEEE, 2017, pp. 736–744.

[29] H. Fan, L. Zheng, and Y. Yang, "Unsupervised person re-identification: clustering and fine-tuning," *ACM Transactions on Multimedia Computing, Communications and Applications*, 2018.

[30] L. Zhang, B. Ma, G. Li, Q. Huang, and Q. Tian, "Generalized semi-supervised and structured subspace learning for cross-modal retrieval," *IEEE Transactions on Multimedia*, vol. 20, no. 1, pp. 128–141, 2018.

[31] R. Rehurek and P. Sojka, "Software framework for topic modelling with large corpora," in *In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. Citeseer, 2010.

[32] M. D. Lee, B. Pincombe, and M. Welsh, "An empirical evaluation of models of text document similarity," in *Proceedings of the Annual Meeting of the Cognitive Science Society*, vol. 27, no. 27, 2005.