Multimodal Joint Representation for User Interest Analysis on Content Curation Social Networks

Lifang Wu, Dai Zhang, Meng Jian(✉), Bowen Yang, and Haiying Liu

Faculty of Information Technology, Beijing University of Technology, Beijing, China
jianmeng648@163.com

Abstract. Content curation social networks (CCSNs), where users share interests by images and their text descriptions, are booming social networks. For the purpose of fully utilizing user-generated contents to analysis user interests on CCSNs, we propose a framework of learning multimodal joint representations of pins for user interest analysis. First, images are automatically annotated with category distributions, which benefit from the network characteristics and represent interests of users. Further, image representations are extracted from an intermediate layer of a fine-tuned multilabel convolutional neural network (CNN) and text representations are obtained with a trained Word2Vec. Finally, a multimodal deep Boltzmann machine (DBM) are trained to fuse two modalities. Experiments on a dataset from Huaban demonstrate that using category distributions instead of single categories as labels to fine-tune CNN significantly improve the performance of image representation, and multimodal joint representations perform better than either of unimodal representations.

Keywords: Multimodal · Content curation social networks
User modeling · Recommender systems

1 Introduction

Content curation social networks (CCSNs) are interest-driven social networks where users can organize and demonstrate multimedia contents they like. Since the most typical CCSN Pinterest became the fastest social network to reach 10M users [4], CCSNs have become popular worldwide. In China, more than 50 Pinterest-like websites such as Huaban, Duitang, Meilishuo, Mogujie and so forth have been published. The rapid development of CCSNs attracts much attention on different research topics, for example, network characteristic analysis [4], user

Supported by National Natural Science Foundation of China 61702022, Beijing Municipal Education Commission Science and Technology Innovation Project KZ201610005012, China Postdoctoral Science Foundation funded project 2017M610027 and 2018T110019.

J.-H. Lai et al. (Eds.): PRCV 2018, LNCS 11258, pp. 363–374, 2018.
https://doi.org/10.1007/978-3-030-03338-5_31

behavior study [5], influence analysis [18], search engine [21], recommender systems [2,9,10,19] and user modeling [1,3,20].

On CCSNs, the carrier of user interests is the basic unit of the network called "pin", which comprises an image and its text description. Most prior works on CCSNs only focused on unimodal data. Yang et al. [19] modeled boards with text representations and recommended boards re-ranked with image representations. Cinar et al. [1] separately predicted categories of pins with either image representations or text representations and fused the results of two modalities by decision fusion. Liu et al. [10] used unimodal representations to respectively generate candidate pins and to re-rank all the candidates. All these methods are late fusion methods which cannot obtain multimodal joint representations.

Multimodal joint representation commonly consists of unimodal representation and multimodal fusion. With regard to image representation, convolutional neural networks (CNNs) have recently achieved many outstanding performances on computer vision. Some works have been done on employing CNNs to represent pins. A key to train CNNs is to create a large labelled dataset. Cinar et al. [1], and You et al. [20] directly used the category of a pin as its label, but this label may be inaccurate as different users may select different categories for a same image. Geng et al. [3] constructed an ontology in fashion domain and trained a multi-task CNN with concepts in ontology, but this methods is hard to be deployed in all domains. Zhai et al. [21] obtained more detailed labels by taking top text search queries on Pinterest, however, the quality and consumption of this annotation highly depends on the performance of the search engine. Inspired by the fact that categories predefined by CCSNs are not independent objects but related notions, we use category distributions based on statistics as labels and fine-tuned a multilabel CNN for image representation.

Many multimodal fusion studies have been carried out on classification and retrieval. Most existed methods are based on discriminative models such as latent Dirichlet allocation [15], CNN [11] and recurrent neural network [12]. Those methods mainly learn the consistency between modalities and can hardly deal with missing input modalities. On the generative side, restricted Bolzmann machine (RBM) [6], deep autoencoder (DAE) [14] and deep Boltzmann machine (DBM) [17] are proved to be feasible to learn both the consistency and complementarity between modalities and can easily deal with the absence of some modalities, however, limited works have been done on fusing features obtained by deep learning with these models. Zhang et al. [22] fused visual features extracted from the 6-th layer of AlexNet and textual features generated by sparse coding of word vectors from a Word2Vec [13] with a DAE. Since DAEs are deterministic models while DBMs are probabilistic models, we trained a multimodal DBM to improve generalization performance.

The proposed framework of learning multimodal joint representations of pins is shown in Fig. 1. For image representation, visual features are extracted from an intermediate layer of the fined-tuned CNN. For text representation, distributed representations of words are learned on corpora and are encoded to represent texts. As our choice, Word2Vec is a frequently used distributed representation for capturing semantic and syntactic relations between words. Mean vector [8] of Word2Vec performs well on text representation and is unsupervised. Our multimodal joint representations is finally generated by a pretrained modified multimodal DBM.

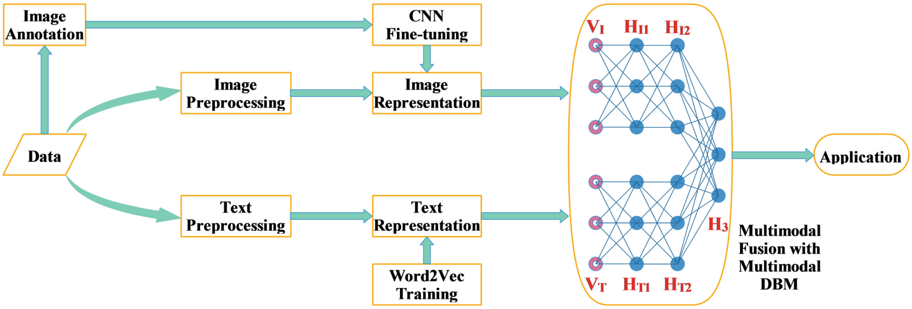


Fig. 1. Framework of learning multimodal joint representations of pins.

We believe that our research is the foundation of further researches on CCSNs such as board and user modeling, with the following contributions:

- We propose an easy-to-accomplish automatic annotate method that accumulate category selections of users to form category distributions of pins and fine-tune a multilabel CNN which significantly improves the category prediction performance.
- Multimodal joint representations of pins we get performs better than the unimodal representations.

The rest of the paper is organized as follows. Section 2 describes the proposed framework in details. Experiments and the corresponding analysis are provided in Sect. 3. And it is followed by conclusions in Sect. 4.

2 Multimodal Joint Representations of Pins

A pin comprise an image and its text description. As shown in Fig. 1, the whole process of multimodal joint representation can be roughly divided into three parts: image representation, text representation and multimodal fusion.

2.1 Image Representation

The aim of image representation is to learn features which not only maintain intrinsic characteristics of images but also relate to user interests on CCSNs. As supervised learning models, CNNs can certainly capture the relationships between images and user interests if user interests on CCSNs are used as labels during the learning process. Not to mention that top layers of CNNs can learn high-level image features, which can be interpreted as color, material, texture, object, scene and so on by some means.

All pins on CCSNs are collected into boards. When a board is created, the owner must select one of categories predefined by CCSNs for it, and all pins in this board will have the same category as the board. Since the category can be considered as the theme of the board, it can be directly treated as a label, which describes a coarse-grained user interest. However, this label is probably weak and noisy, mainly because user preferences may lead to various category selections for a same image since it can be observed that categories in Table 1 are sometimes related notions. To put it in practical terms, the image in Fig. 2a may belong to photography, kids and pets on Huaban.

Table 1. List of all 33 predefined categories on Huaban.

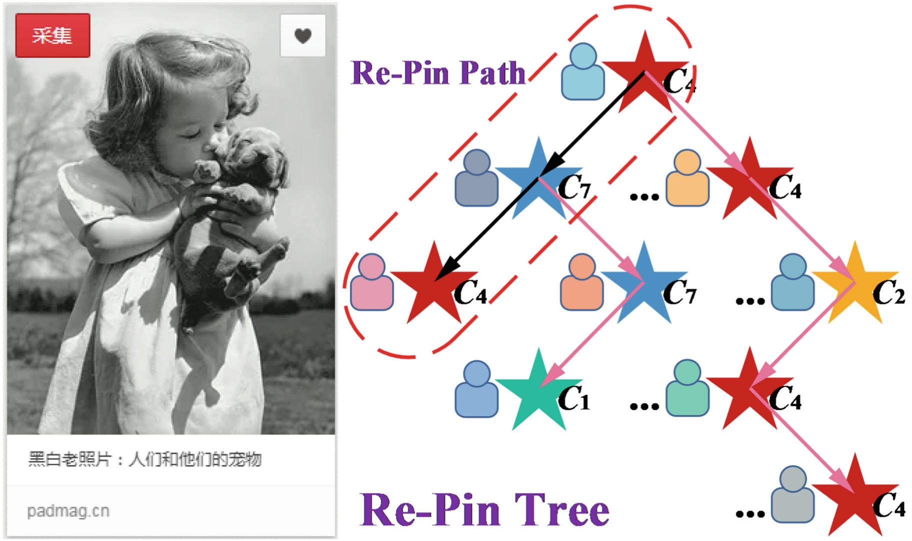
Anime	Apparel	Architecture	Art	Beauty
Cars motorcycles	Data presentation	Design	Desire	DIY crafts
Education	Film music books	Fitness	Food drink	Funny
Games	Geek	Home	Illustration	Industrial design
Kids	Men	Modeling hair	People	Pets
Photography	Quotes	Sports	Tips	Travel places
Web app icon	Wedding events	**Other**		

The most frequently activity on CCSNs is called “re-pin”, which means a user collects an image and may add a text description for it from a board of another user into his or hers own board. A “re-pin path” is formed if users are interested in a same image and thus they re-pin it one by one, and all re-pin paths of an image form a “re-pin tree”, as illustrated in Fig. 2b. Because any one of the categories in the re-pin tree cannot decide what this image is about but describes a portion of it instead, we use a category distribution to represent interests of an image. The category distribution of a given image I can be computed after counting the categories in the re-pin tree as

$$Interest_I = \left(p_{C_i} = \frac{f_{C_i}}{\sum\limits_{i=1}^{N_C} f_{C_i}} \right) \in [0,1]^{N_C} \tag{1}$$

where f_{C_i} denotes the frequency of the i-th category C_i, N_C is the total number of categories on CCSNs. In practice we set

$$f_{C_i} = 0 \quad \text{if} \quad f_{C_i} < \frac{\sum\limits_{i=1}^{M_C} f_{C_i}}{M_C} \tag{2}$$



(a) (b)

Fig. 2. (a): Example of a pin on Huaban. (b): Illustration of a re-pin tree composed of some re-pin paths. Each star represents a pin and C_i nearby is the category of this pin. All pins in the re-pin tree have a same image.

where M_C is the total number of categories occurred in the re-pin tree to filter out spam and make the sequence on behalf of majority opinion.

After automatic image annotation, we then choose a pretrained CNN model to fine-tune for the purpose of accelerate the training process. Most available pretrained CNNs are designed for classifying independent objects, while our model should be a multilabel regressor. Accordingly, we change the loss layer from softmax with logarithmic loss layer to sigmoid with cross entropy loss layer. The loss function is defined as

$$\mathrm{E} = -\sum_{i=1}^{N_C} \left[p_{C_i} \ln \hat{p}_{C_i} + (1 - p_{C_i}) \ln (1 - \hat{p}_{C_i}) \right] \tag{3}$$

where p_{C_j} is the percentage in Eq. (1), $\hat{p}_{C_j}$ denotes the corresponding sigmoid output.

After fine-tuning, the weights of the CNN are stored for feature extraction. The activation values of an fully connected (FC) layer will be extracted as the image representations.

2.2 Text Representation

An important aim of text representation is also to discover the relationships between descriptions of pins and categories. However, it is difficult to create a large labelled dataset on CCSNs for supervised learning as descriptions in the re-pin tree may be different.

Since there is no obvious difference between words used on CCSNs and those in common situations, we train a Word2Vec on some public corpora to encode words. Word2Vec is an efficient shallow model for learning distributed representations of words. Although the learning process of either of its two log-linear models, which are continuous bag-of-words (CBOW) and continuous skip-gram, is supervised, there is no need to annotate the training texts. Since the learned vectors capture a large number of meaningful semantic and syntactic word relationships, we make sure that the categories are in the training dictionary in order that the relationships between words and the categories can be considered as the relationships between words and user interests. In addition, distributed representations are scalable even though the vocabulary of natural language is extremely wide.

Owing to the fact that texts have diverse lengths, it is necessary to transform a set of word vectors into a single vector with a constant dimension for representing a complete text. For a text T, the text representation is the mean vector computed as

$$V_T = \frac{1}{M_T} \sum_{i=1}^{M_T} KeyedVector_{Word_i} \tag{4}$$

where $KeyedVector_{Word_i}$ denotes the word vector of the i-th word $Word_i$, M_T is the text length.

2.3 Multimodal Fusion

Different modalities typically have different statistical properties, which makes it difficult to learn a joint representation that capture both consistent and complementary relationships across modalities. A multimodal DBM which combines DBMs by adding a shared hidden layer on top of them can effectively solve this problem. A DBM is structured by stacking RBMs in a hierarchical manner. A RBM is an undirected graphical model with binary-valued visible layer V and binary-valued hidden layer H fully connected to each other defines the energy function

$$\mathrm{E}\left(V, H; \theta\right) = -H^T W V - A^T V - B^T H \tag{5}$$

where $\theta = \{W, A, B\}$ denotes the model parameters including the symmetric interaction terms W between two layers, visible layer bias terms A and hidden layer bias terms B.

As illustrated in Fig. 1, we use two-layer DBMs with Gaussian-Bernoulli RBMs, which are a variant of RBMs that can model real-valued vectors, as bottom for both modalities. A Gaussian-Bernoulli RBM with visible units $V = \{v_i\} \in \mathbb{R}^D$ and hidden units $H = \{h_j\} \in \{0,1\}^F$ defines the energy function

$$\mathrm{E}\left(V, H; \theta\right) = \sum_{i=1}^{D} \frac{\left(v_i - a_i\right)^2}{2{\sigma_i}^2} - \sum_{i=1}^{D}\sum_{j=1}^{F} \frac{v_i}{\sigma_i} w_{ij} h_j - \sum_{j=1}^{F} b_j h_j \tag{6}$$

where σ_i denotes the standard deviation of the i-th visible unit and $\theta = \{\{w_{ij}\} \in \mathrm{I\!R}^{D\times F}\}, \{a_i\} \in \mathrm{I\!R}^{D}\}, \{b_j\} \in \mathrm{I\!R}^{F}\}, (\sigma_i) \in \mathrm{I\!R}^{D}\}$. During the unsupervised training of the multimodal DBM, modalities can be thought of labels for each other. Since RBMs can be considered as autoencoders, each layer of the multimodal DBM makes a small contribution to eliminate modality-specific correlations. Consequently, the top layer can learn a relatively modality-free representation as opposed to the modality-full input layers. The joint distribution over the multimodal inputs can be written as

$$\mathrm{P}\left(V_I, V_T; \theta\right) = \sum_{H_{I2}, H_{T2}, H_3} \mathrm{P}\left(H_{I2}, H_{T2}, H_3\right) \left(\sum_{H_{I1}} \mathrm{P}\left(V_I, H_{I1}, H_{I2}\right)\right) \left(\sum_{H_{T1}} \mathrm{P}\left(V_T, H_{T1}, H_{T2}\right)\right) \quad (7)$$

where θ denotes all model parameters.

A pin may has no text description. The multimodal DBM can be used to generate missing text representation by sampling it from the conditional distribution with the standard Gibbs sampler. Finally, activation probabilities of H_3 are used as the multimodal joint representations of pins no matter they have text descriptions or not.

3 Experiment

3.1 Dataset and Implementation Details

All data used in experiments was crawled from Huaban, which is one of typical CCSNs in China. Huaban provides almost the same applications as Pinterest provides, while three main differences between them are: users can "like" pins or boards on Huaban while "like" has been removed by Pinterest; Huaban records both users from whom a pin re-pinned and by whom it initially created while Pinterest only records the direct source; some predefined categories are different and Pinterest has 5 more categories.

We first crawled pins without images of 5957 users and sampled 88 users according to pin counts and categories of their boards. To make our dataset diverse and real, a few cold start and extremely active users have been confirmed in it. We then downloaded the images of sampled users and pins of their like boards. In addition, top 1000 recommended pins of every category was crawled for fine-tuning the CNN, and re-pin paths of all recommended pins was crawled for automatic annotation. In total, the dataset includes 1694 boards and 167747 unique images. All pins was used as supplements for obtaining category distributions of all recommended pins. The average nodes of the incomplete re-pin trees is 47.57.

Labeled images was split into 80% for training and validating and the remaining 20% for testing after label balancing. AlexNet [7] with ImageNet [16] pre-trained weights was chosen as a basis. Because AlexNet requires a constant input dimension, the image was first rescaled such that the shorter side was of length 256 pixels, and then the central 256 × 256 patch of the resulting image was cropped out. As a comparison, we also used the most frequent category as label to fine-tune an multiclass AlexNet. The dimensions of the fc8 layers of both CNNs were change to 33. Image representations was extracted from the FC7 layer of the fine-tuned CNN.

Word2Vec was trained on Wikipedia dumps and Sougou Lab dataset with CBOW and negative sampling. The vector dimension was set to 300. Words with total frequency lower than 5 are ignored. Preprocessing such as traditional Chinese and simplified Chinese conversion, removing punctuation, word tokenize, removing stop words and machine translation has been done on text descriptions of pins.

Image and text features were used for pretraining our multimodal DBM. Dimensions of H_{T1}, H_{T2} and H_{V1} were equal to their corresponding visible inputs, and dimension of H_{V2} and H_3 was set to 2048 for the purpose of compressing the vectors. DBM was pretrained using a greedy layer-wise strategy by learning a stack of modified RBMs. Finally, we ran Gibbs sampler to generate missing text representations and to infer multimodal join representations.

3.2 Analysis of Interests Represented by Pins

Analysis of interests represented by pins is the prerequisite of analysis of interests represented by boards and user interest analysis. The category distribution are interests of the image and can be approximate the interests of the pin, even though some of categories will be enhanced by the text description.

Table 2. Comparison on pin category prediction

Model	Dimension	Dominant category accuracy	Mean nonzero error	Mean error
AlexNet [1]	4096	57.53%	—	—
Word2Vec [1]	300	33.47%	—	—
AlexNet [20]	4096	43.1%	—	—
AlexNet	4096	45.85%	—	—
Multilabel	4096	82.71%	0.1320	0.0141
Word2Vec	300	42.88%	0.3249	0.0415
Multimodal	2048	84.13%	0.1181	0.0119

Multidimensional logical regressions (LRs) were trained on recommended pins for all unimodal and multimodal representations. The results are shown in Table 2, together with the result of the compared AlexNet. Relevant results on 32 [1] and 34 [20] Pinterest categories are also cited as references. Mean nonzero error is the average error between all nonzero categories and corresponding predictions. The dominant category accuracy checks the consistency of the most frequently category between predictions and labels. Comparision of two fine-tuned CNNs shows that our multilabel regressor significantly improves the accuracy.

It is because that category distributions can not only eliminate the interference of related categories but also provide more information to learn than only dominant categories. Although the performance of text representations is not comparable with those of image representations, the complementarity between two modalities helps the multimodal joint representations perform better than the unimodal representations. Our framework can also infer interests of images from other social networks.

3.3 Board Category Recommendation

Every board must be assigned a category nowadays, while some boards have no category as a result of that they were created before the constraint entered into force. However, it is illogical because even if it is hard to select a category for a board about wide interests, "other" in Table 1 can be selected. Consequently, Huaban offers a function that allows any user to select a category for a board which haven't categorized. Board category recommendation will be useful on that occasion, and the first selection and further editing too.

Table 3. Comparison on board category recommendation

Model	Top-1 MRR	MRR
Random	3.03%	12.39%
Text + Cosine similarity	25.65%	38.78%
Image + Multidimensional LR	60.10%	73.41%
Text + Multidimensional LR	38.00%	54.30%
Multimodal + Multidimensional LR	62.35%	74.77%

Same as interests represented by pins, interests represented by boards should not limited in one category. The interest distribution of a board can be computed by averaging all category distributions of its pins. As pins are accumulated, the category preference is reinforced due to the fact that the accumulation process of strong categories are faster than those of weak categories. Our recommended category is the max category in the interest distribution of the board, and the ground truth is the real category of the board. Mean reciprocal rank (MRR) are used as the performance metric. As board category recommendation actually has only one correct selection, we also give the top-1 MRR. Results are organized in Table 3. Cosine similarities between texts and categories are less effective than category distributions obtained with texts, this indicates that there is a gap between semantemes and interests. The multimodal joint representations, which benefit from personalized texts, perform better than image representations. Notice that the recommendation dataset is different from the training dataset, it also proves that our framework has a good generalization ability.

The first selection is a cold start problem, as it only depends on one pin. We then evaluate the influence of pin counts on board category recommendation.

Table 4. Influence of pin count on board category recommendation based on multimodal join representation

Pin count	Top-1 MRR	MRR
=1	7.69%	35.56%
≤4	45.57%	59.68%
≤30	56.53%	69.75%
≤100	59.06%	72.39%
>100	67.11%	78.33%

As shown in Table 4, our recommendation suffers the cold start. However, the theory about preference reinforcement is proved as more pins lead to better performance. Although interests of users are more discrete than interests of boards, we infer that the accumulation process is still effective on user interest analysis.

3.4 Board Recommendation

Well organized boards can be high quality galleries, which makes it easier for users to collect pins. For this reason, CCSNs offer users a board recommendation function. Besides interest distributions, a board can be represent by the mean vector of representations of pins. And similarity between boards can be simply measured with some distance metrics, for example cosine similarity.

Table 5. Comparison on board recommendation

Model	Top-5 MRR	MRR
Category based	2.12%	3.85%
Image + Multidimensional LR	16.08%	18.61%
Text + Multidimensional LR	15.93%	17.95%
Multimodal + Multidimensional LR	17.58%	20.13%
Image + Mean vector	33.66%	35.97%
Text + Mean vector	25.96%	27.49%
Multimodal + Mean vector	35.76%	37.88%

We divided every board in half according to the order of pins, and each half must be similar board for another. The owner of each half will be interested in another half and further re-pin from or like or follow it beyond all doubt. On the basis of this, we consider half of the board as the only correct recommendation result and retrieve the index in the similarity sequence. As Huaban exhibit five pins at the top row of its waterfall flow for common resolution screens, we also demonstrate top-5 MRR. Table 5 shows that results of mean vectors is higher than those of respective interest distributions, simply owing to the additional information. All of our methods significantly improve the results in comparison with the category based filtering. The results also show that multimodal joint representations can model boards better than either of unimodal representations.

4 Conclusion

We propose a framework of learning multimodal joint representations of pins on CCSNs. Experimental results show that multimodal joint representations performs better than either of unimodal representations on interpreting pin-level interests and board-level interests. The obtained representations can be easily used on user modeling and recommender systems for CCSNs. Future work will be focused on extending our framework to model boards and users. In addition, other effective feature extraction methods and multimodal fusion approaches may be taken into account.

References

1. Cinar, Y., Zoghbi, S., Moens, M.F.: Inferring user interests on social media from text and images. In: 2015 IEEE International Conference on Data Mining Workshop (ICDMW), pp. 1342–1347. IEEE (2015)
2. Geng, X., Zhang, H., Bian, J., Chua, T.: Learning image and user features for recommendation in social networks. In: Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), pp. 4274–4282. IEEE Computer Society (2015)
3. Geng, X., Zhang, H., Song, Z., Yang, Y., Luan, H., Chua, T.: One of a kind: user profiling by social curation. In: Proceedings of the 22nd ACM International Conference on Multimedia, pp. 567–576. ACM (2014)
4. Gilbert, E., Bakhshi, S., Chang, S., Terveen, L.: "i need to try this!": a statistical overview of pinterest. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 2427–2436. ACM (2013)
5. Han, J., et al.: Sharing topics in pinterest: understanding content creation and diffusion behaviors. In: Proceedings of the 2015 ACM on Conference on Online Social Networks, pp. 245–255. ACM (2015)
6. Jia, X., Wang, A., Li, X., Xun, G., Xu, W., Zhang, A.: Multi-modal learning for video recommendation based on mobile application usage. In: 2015 IEEE International Conference on Big Data (Big Data) (BIG DATA), pp. 837–842. IEEE (2015)
7. Krizhevsky, A., Sutskever, I., Hinton, G.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems 25, pp. 1097–1105. Curran Associates Inc. (2012)
8. Lev, G., Klein, B., Wolf, L.: In defense of word embedding for generic text representation. In: Biemann, C., Handschuh, S., Freitas, A., Meziane, F., Métais, E. (eds.) NLDB 2015. LNCS, vol. 9103, pp. 35–50. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-19581-0_3
9. Li, Y., Mei, T., Cong, Y., Luo, J.: User-curated image collections: modeling and recommendation. In: 2015 IEEE International Conference on Big Data (Big Data), pp. 591–600. IEEE (2015)

10. Liu, D., et al.: Related pins at pinterest: the evolution of a real-world recommender system. In: Proceedings of the 26th International Conference on World Wide Web Companion, pp. 583–592. International World Wide Web Conferences Steering Committee (2017)
11. Ma, L., Lu, Z., Shang, L., Li, H.: Multimodal convolutional neural networks for matching image and sentence. In: 2015 IEEE International Conference on Computer Vision (ICCV) (2015), pp. 2623–2631. IEEE (2015)
12. Mao, J., Xu, J., Jing, Y., Yuille, A.: Training and evaluating multimodal word embeddings with large-scale web annotated images. In: Advances in Neural Information Processing Systems 29, pp. 442–450. Curran Associates, Inc. (2016)
13. Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Proceedings of the 26th International Conference on Neural Information Processing Systems, vol. 2, pp. 3111–3119. Curran Associates Inc. (2013)
14. Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., Ng, A.: Multimodal deep learning. In: Proceedings of the 28th International Conference on Machine Learning, pp. 529–545. Omnipress (2011)
15. Qian, S., Zhang, T., Xu, C.: Multi-modal multi-view topic-opinion mining for social event analysis. In: Proceedings of the 2016 ACM on Multimedia Conference, pp. 2–11. ACM (2016)
16. Russakovsky, O., Salakhutdinov, R.: Imagenet large scale visual recognition challenge. Int. J. Comput. Vis. **115**, 211–252 (2015)
17. Srivastava, N., Salakhutdinov, R.: Multimodal learning with deep boltzmann machines. J. Mach. Learn. Res. **15**, 2949–2980 (2014)
18. Venkatadri, G., Goga, O., Zhong, C., Viswanath, B., Gummadi, K., Sastry, N.: Strengthening weak identities through inter-domain trust transfer. In: Proceedings of the 25th International Conference on World Wide Web, pp. 1249–1259. ACM (2016)
19. Yang, X., Li, Y., Luo, J.: Pinterest board recommendation for twitter users. In: Proceedings of the 23rd ACM International Conference on Multimedia, pp. 963–966. ACM (2015)
20. You, Q., Bhatia, S., Luo, J.: A picture tells a thousand words-about you! user interest profiling from user generated visual content. Signal Process. **124**, 45–53 (2016)
21. Zhai, A., et al.: Visual discovery at pinterest. In: Proceedings of the 26th International Conference on World Wide Web Companion, pp. 515–524. International World Wide Web Conferences Steering Committee (2017)
22. Zhang, H., Yang, Y., Luan, H., Yan, S., Chua, T.: Start from scratch: towards automatically identifying, modeling, and naming visual attributes. In: Proceedings of the 22nd ACM International Conference on Multimedia, pp. 187–196. ACM (2014)