Personalized Recommendation of Tumblr Posts Using Graph Convolutional Networks with Preference-aware Multimodal Features

Kazuma Ohtomo †, Ryosuke Harakawa (member) †, Takahiro Ogawa (member) ††, Miki Haseyama (fellow) ††, Masahiro Iwahashi (member) †

Abstract
Tumblr is a popular micro-blogging service on which users can share posts comprising text and images. This paper presents a method for personalizing post recommendations for each user from a large number of posts. Specifically, we develop a supervised multi-variational auto encoder considering user preference (SMV AE-UP). SMV AE-UP can extract relationships between text and image features by considering class information representing a user’s preference for each post; thus, preference-aware multimodal features can be calculated. Furthermore, for each target user, a network that enables comparison between a user and posts in the same feature space is constructed using the preference-aware multimodal features and metadata on posts. By applying graph convolutional networks (GCNs) to the network constructed for each target user, an accurate recommendation matching each user’s preferred posts becomes feasible. Experimental results for real-world datasets including six users and 99,844 posts show the effectiveness of our method.

Key words: Tumblr, recommendation, multi-variational auto encoder (MVAE), graph convolutional networks (GCN).

1. Introduction

Tumblr∗ is one of the most popular micro-blogging services with over 400 million users. On Tumblr, items of content uploaded by users are called posts. Tumblr has functions such as Like that lets users indicate their post preferences and Reblog that lets them re-share posts by other users. In Tumblr, the number of posts is too large for each user to easily find posts of interest. Therefore, it is necessary to develop a method for accurately recommending posts of interest to each user.

A typical Tumblr post comprises text and/or images [1]. As shown in Fig. 1, we can also find the network structure of users using “following,” “Like” or “Reblog” relations, which collectively reflect each user’s preferences. Therefore, we propose using the network structure as well as text and image features to generate post recommendations that better reflect each user’s preferred post types.

Many methods have been proposed for recommending Web content using multiple kinds of features [2–6].

Fig. 1: Illustration of the network structure on Tumblr∗∗.

Previous works [2–4] enabled recommendation by extracting latent features representing the semantics of Web content from multiple features using multimodal deep restricted Boltzmann machine (mmRBM) [2], cross media latent dirichlet allocation (LDA) [3] and multimodal field-aware factorization machines (FFM) [4]. Recently, researchers have proposed methods that analyze the network structure using Web content metadata and content features [5, 6]. Because metadata such as favorite and uploader information reflect each user’s preferences, the recommendation performance can be improved by analyzing the network structure.

Inspired by the aforementioned previous works [2–6], we propose a novel method for recommending each user’s desired Tumblr post types using a network structure that reflects user preferences as well as text and image features. There are the following two reasons why we target Tumblr in this paper.

• To the best of our knowledge, few studies on multimodal analysis for Tumblr have been done. Thus, our work is valu-
able as the first report that targets Tumblr.

- Because Tumblr has a larger number of users than other popular services such as Twitter and Pinterest, our work that targets Tumblr is valuable as a real-world application. Specifically, Tumblr has 492 million users while Twitter and Pinterest have 330 and 367 million users respectively.

First, we derive a novel multimodal feature extraction method called supervised multi-variational auto encoder considering user preference (SMVAE-UP). SMVAE-UP can extract relationships between text and image features by considering user preferences; thus, preference-aware multimodal features can be calculated. Specifically, we introduce triplet loss [7], which includes class information that represents whether a user prefers each post in a multi-variational autoencoder (MVAE) [8]. This is one of the most successful methods for latent feature extraction. Thus, we can derive a SMVAE-UP that can extract preference-aware multimodal features reflecting the user’s preferences and semantics of posts. Furthermore, for each target user, we construct a network that enables comparison between the user and posts in the same feature space using the obtained preference-aware multimodal features and metadata of Reblog. Then, with the network constructed for each user, we classify whether a user prefers each post using graph convolutional networks (GCN) [9], one of the most successful classifiers. This approach enables more accurate post recommendations for each user. Experimental results for real-world datasets involving six users and 99,844 posts show the effectiveness of our proposed approach, i.e., latent feature extraction based on SMVAE-UP and recommendations based on GCN.

2. Related Work

To clarify the novelty and contribution of this work, we present related works addressing latent feature extraction from multiple feature (Sec. 2.1) and recommendation types using both content features and the network structure (Sec. 2.2).

2.1 Latent Feature Extraction from Multiple Feature Types

Many methods have been proposed to project multiple feature types into the same feature space. These include canonical correlation analysis (CCA) [10], partial least squares (PLS) and cross-modal factor analysis (CFA) [11]. These methods calculate linear transformations that maximize correlations among multiple feature types. However, these methods are unsuitable for very large content datasets because we must solve the eigenvalue problem. Instead, LDA-based or deep learning-based methods have recently been proposed [8,12–17]. LDA is one of the most representative topic models and extended versions that handle multiple types of data (e.g., multi-modal LDA (mMLDA) [12] and correspondence LDA [13]) have been proposed. However, extended models [14,15] of restricted Boltzmann machine (RBM) [18] and auto encoder (AE) [19] have also been proposed and which enable extraction of discriminative latent features from multiple types of data. Furthermore, extended models of variational auto encoder (VAE) [20], an unsupervised deep learning technique that can be applied to very large content datasets, have been proposed. Specifically, Suzuki et al. [16] presented a model that projects multiple features into the same latent space by concatenating multiple encoders. Other papers [8,17] have proposed models that calculate latent features by integrating probability distributions estimated by multiple encoders to the same distribution via Product of Expert (PoE) [21]. Additionally, Mike et al. [8] propose a weakly supervised MVAE. The weakly supervised MVAE uses class information as one of the input features; thus, we can successfully extract latent features that generate both the input original features’ class information.

Compared with the above methods, our proposed SMVAE-UP is novel in that it learns the latent features using triplet loss [7] with class information so that distances between the same (different) class samples can be close (distant). Thus, we can successfully calculate preference-aware multimodal features.

2.2 Recommendation Based on Collaborative Use of Content Features and Network Structure

Recently, many papers (e.g., [5,22–25]) have reported that the combined use of content features and network structure improved the performance of Web content analysis. For a typical task in this domain, i.e., recommendation, the effectiveness of using both content features and the network structure has been confirmed [5,23,24]. For example, Ying et al. [23] proposed a method for recommending Pinterest contents by applying GCN to a network in which image features are embedded. Matsumoto et al. [5] proposed a method for recommending music videos on the basis of link prediction on a network that represented relationships between users and contents via sub-sampled CCA [25] based latent feature extraction.

---

**Notes:**

Inspired by these studies [22–25], we construct a network that represents relationships between users and posts with consideration of each user’s preferences based on SMVÆ-UP. Then we realize recommendation of each user’s preferred posts by applying GCN [9] to the constructed network.

3. Recommendation of Tumblr Posts Based on GCN with SMVÆ-UP

In this section, we explain the details of the proposed method, including the preference-aware multimodal feature extraction via SMVÆ-UP (Sec. 3.1) and recommendations based on GCN (Sec. 3.2).

3.1 Latent Feature Extraction Based on SMVÆ-UP

We derive the SMVÆ-UP by introducing a triplet loss [7] into the MVAE [8]. The model architecture of the SMVÆ-UP is illustrated in Fig. 2. We formulate each user’s preferences as the triplet loss using the metadata item Like and text and image features (defined below) of Tumblr posts. In the experiment, we assigned class 1 (class 0) to posts Liked (not liked) by the target user. To approach posts of class 1 and separate posts of class 0, we formulate the triplet loss \( L_{\text{Triplet}} \) as:

\[
L_{\text{Triplet}} = \gamma \cdot \max(0, z^T z_n - z^T z_p + m).
\]

Here, \( z, z_p, z_n, \gamma \) and \( m \) are the target multimodal features, features belonging to the same class, features belonging to different classes, a hyperparameter, and a margin, respectively. Using the MVAE [8], we calculate text and image features in each post and obtain probability distributions using multiple encoders. The recently presented papers [26,27] report that term-based features such as TF-IDF have more discriminative power than semantic embedding-based features such as word2vec [28] and doc2vec [29] in the classification task on Twitter. Other recent studies [4,22] also show that TF-IDF is useful with Twitter data for clustering and recommendation. Texts on Tumblr and Twitter have similar characteristics in that the text often includes abbreviations, emoticons, acronyms and Internet slang words. Therefore, we adopted TF-IDF in the experiment. However, the TF-IDF features have too many dimensions to efficiently train the SMVÆ-UP. Thus, we applied principal component analysis (PCA) to the TF-IDF features and obtained 128-dimensional features for use as the text features \( x_{\text{text}} \) in our experiment. Moreover, in line with recently presented papers (e.g. [30,31]), we used VGG16 [32] for social media analysis. We calculated 25088-dimensional image features \( x_{\text{image}} \) from the fifth pooling layer of VGG16 pretrained on ImageNet [33]. In the same manner as text feature extraction, we obtained 128-dimensional features using PCA.

To implement the SMVÆ-UP, we need to calculate \( p_\theta(z|x) \). However, because \( p_\theta(z|x) \) is a posterior distribution including an unknown variable \( z \), it is difficult to calculate directly.

To overcome this difficulty, we prepare \( q_\phi(z|x) \), which approximates \( p_\theta(z|x) \). To realize this approximation, we need to minimize Kullback–Leibler divergence between the distributions of an encoder \( q_\phi \) and a decoder \( p_\theta \) as follows:

\[
\arg \min_{\phi, \theta} \text{KLD}(q_\phi(z|x)||p_\theta(z|x)).
\] (2)

Here, \( x \in \{ x_{\text{text}}, x_{\text{image}} \} \) denotes the input features, and \( z \) represents the multimodal features. Following [34], we rewrite the divergence as:

\[
\text{KLD}(q_\phi(z|x)||p_\theta(z|x)) = \log p_\theta(x) - \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] - \text{KLD}(q_\phi(z|x)||p_\theta(x)).
\] (3)

Here, the second term on the right side of this equation is called the evidence lower bound (ELBO). In this study, we propose a modified ELBO (MELBO) so that we can obtain \( z \) that reflects each user’s preferences. Specifically, we incorporate the triplet loss from Eq. (1) into the conventional ELBO as follows:

\[
\text{MELBO}(x) = \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] - \text{KLD}(q_\phi(z|x)||p_\theta(x)) - L_{\text{Triplet}}.
\] (4)

Therefore, we can rewrite Eq. (3) as:

\[
\text{KLD}(q_\phi(z|x)||p_\theta(z|x)) = \log p_\theta(x) - \text{MELBO}(x).
\]

By maximizing the MELBO, we can obtain \( q_\phi(z|x) \) that approximates \( p_\theta(z|x) \). Next, we further modify the MELBO to calculate \( z \) that reflects the multimodal features of Tumblr posts. To this aim, for the set of multimodal features
focus only on the semantics of the items. (8) The term \(\sum_{j \in V} E_{ij} \cdot \log \rho(y_i|z)\) represents the classification loss, which is used to predict the class of each input. The term \(-\beta \cdot \text{KL}(p_{\phi}(z|x)||p_{\phi}(z))\) is a regularization term that encourages the learned posterior distribution \(p_{\phi}(z|x)\) to be close to the prior distribution \(p_{\phi}(z)\). The term \(-L_{\text{Triplet}}\) is a triplet loss term that encourages the learned posterior distribution \(p_{\phi}(z|x)\) to be close to the prior distribution \(p_{\phi}(z)\). The term \(L_{\text{graph}}\) is a graph loss term that encourages the learned posterior distribution \(p_{\phi}(z|x)\) to be close to the prior distribution \(p_{\phi}(z)\). The term \(h_{i+1} = \rho(\sum_{r \in E} \sum_{j \in N_r(i)} 1 W_r W_0') + W_l h_i\) is the graph convolution operation, which is used to update the representation of the target node.

In the experimental setting, posts that are preferred by the target user are defined as preferred posts. We also used the Adam optimizer [35] to train the GCN. By classifying whether each post corresponds to the target user’s preferences, that is, whether each post is Liked, personalized Tumblr post recommendations become feasible.

4. Experimental Results

From Tumblr, we first select a target user. Then we crawled posts that the target user Liked as well as posts uploaded by other users that the target user follows. As described in Section 3.1, we extracted the crawled posts, including both text and images, and assigned class 1 (class 0) to posts that the target user Liked (did not like). Table 1 shows the details of the crawled datasets.

Using 99,844 posts viewed by the six users, we verify the effectiveness of our multimodal analysis algorithms including SMVVAE-UP and GCN. The number of posts is larger than the previous work [8] on multimodal analysis that uses 70,000 contents. In this way, the experiments are enough large-scale to verify the effectiveness of our multimodal analysis algorithms.

4.1 Evaluations

For the quantitative evaluation, we used the F-measure defined by the following equation:
Table 1: Details of datasets.

(a) Details of posts.

<table>
<thead>
<tr>
<th></th>
<th>No. of class 1</th>
<th>No. of class 0</th>
</tr>
</thead>
<tbody>
<tr>
<td>User 1</td>
<td>1,279</td>
<td>4,296</td>
</tr>
<tr>
<td>User 2</td>
<td>722</td>
<td>10,698</td>
</tr>
<tr>
<td>User 3</td>
<td>659</td>
<td>17,363</td>
</tr>
<tr>
<td>User 4</td>
<td>7,183</td>
<td>16,237</td>
</tr>
<tr>
<td>User 5</td>
<td>22,603</td>
<td>8,133</td>
</tr>
<tr>
<td>User 6</td>
<td>1,916</td>
<td>8,755</td>
</tr>
</tbody>
</table>

(b) Details of the network structure.

<table>
<thead>
<tr>
<th></th>
<th>No. of nodes</th>
<th>No. of edges</th>
</tr>
</thead>
<tbody>
<tr>
<td>User 1</td>
<td>217,127</td>
<td>3,222,788</td>
</tr>
<tr>
<td>User 2</td>
<td>390,015</td>
<td>7,203,936</td>
</tr>
<tr>
<td>User 3</td>
<td>868,461</td>
<td>20,456,144</td>
</tr>
<tr>
<td>User 4</td>
<td>768,714</td>
<td>15,081,331</td>
</tr>
<tr>
<td>User 5</td>
<td>401,043</td>
<td>5,803,225</td>
</tr>
<tr>
<td>User 6</td>
<td>469,940</td>
<td>7,028,875</td>
</tr>
</tbody>
</table>

F-measure = $2 \times \frac{P \times R}{P + R}$

$R = \frac{TP}{TP + FN}$

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

Here, TP, FN and FP are the numbers of true positive, false negative, and false positive, respectively. For each dataset, we performed five-fold cross validation. We tested different feature extraction methods and classifiers and calculated F-measures of the recommendation results for each case.

(1) Comparison of Features

We evaluated the effectiveness of our preference-aware multimodal features based on the SMVAE-UP shown in Sec. 3.1 by comparing our method with other feature extraction methods. We trained SMVAE-UP for 500 epochs. Following the idea of warm-up [38], we changed $\gamma$ in Eq. (1) and $\beta$ in Eq. (5) from 0 to 1 over 200 epochs. For the subsequent epochs, we set $\gamma$ and $\beta$ to 1. Additionally, $m$ in Eq. (1) and $A_y$ in Eq. (5) were set to 0 and 1, respectively. We set the initial learning rate to 0.00001 and used Swish [39] for the activation function. Furthermore, we trained GCN for 2,000 epochs. The initial learning rate was set to 0.001. For the evaluation, we used the following comparative methods.

**Text:** This method performs classification by GCN similar to the proposed method but using only text features same as our method.

**Image:** This method performs classification by GCN similar to the proposed method but using only image features same as our method.

**CCA [10]:** This method performs classification by GCN using latent features obtained by CCA [10]. Because of the computational cost, we could not solve the eigenvalue problem for Users 2, 3, 4, 5 and 6. Instead, we used a sub-sampled CCA [25] instead of the original CCA.

**LPCCA [36]:** This method performs classification by GCN using latent features obtained by a locality preserving CCA (LPCCA) [36], an improved version of CCA. We defined the number of neighbors $k'$ used in LPCCA to 15 and 30. In the same way as CCA, we performed sub-sampling [25] for Users 2, 3, 4, 5 and 6.

**MVAE [8]:** This method performs classification by GCN using latent features obtained by MVAE [8]. MVAE was trained using the same text and image features as our proposed method.

**MVAE (with label) [8]:** This method performs classification by GCN using latent features obtained by weakly-supervised version of MVAE [8] with class information. This method used not only text and image features but also class information as the input variables.

Table 2 shows the comparison results. The results indicate that the proposed method enables more accurate recommendation than comparative methods. The results from SMVAE-UP (the proposed method), text and image confirm the effectiveness of using multiple feature types. Furthermore, the effectiveness of supervised learning using class information can be confirmed by comparing the SMVAE-UP, CCA, LPCCA, MVAE and MVAE (with label) methods. In sum, we can quantitatively confirm the effectiveness of our personalized recommendation using GCN with preference-aware multimodal features, which can account for each user’s preferences and the semantics of Tumblr posts. For further discussion, using t-SNE [37], we visualized the features used here. Figure 4 indicates the difficulty of splitting two classes by unsupervised methods (Figs. 4 (a)–(f)) or a weakly-supervised method (Fig. 4 (g)). The figure shows that SMVAE-UP (the proposed method) has more discriminative power than other methods used in this experiment.

(2) Comparison of Classifiers

Here, we evaluate our recommendation method based on GCN shown in Sec. 3.2 by comparing our method with other classifiers. We compared our method with the following classifiers.

**SVM (Support Vector Machines) [40]:** We set the number of classes to two. Also, we used Linear kernel and determined its parameters by grid search [43].

**k-NN (k-Nearest Neighbors) [41]:** We set the number
Table 2: Comparison of feature extraction methods. Columns (a) and (b) denote F-measures for classes 0 and 1, respectively.

<table>
<thead>
<tr>
<th></th>
<th>User 1</th>
<th>User 2</th>
<th>User 3</th>
<th>User 4</th>
<th>User 5</th>
<th>User 6</th>
<th>Ave.</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a)</td>
<td>(b)</td>
<td>(a)</td>
<td>(b)</td>
<td>(a)</td>
<td>(b)</td>
<td>(a)</td>
<td>(b)</td>
</tr>
<tr>
<td>Text</td>
<td>0.948</td>
<td>0.818</td>
<td>0.980</td>
<td>0.634</td>
<td>0.962</td>
<td>0.916</td>
<td>0.970</td>
</tr>
<tr>
<td>Image</td>
<td>0.840</td>
<td>0.590</td>
<td>0.892</td>
<td>0.228</td>
<td>0.876</td>
<td>0.180</td>
<td>0.826</td>
</tr>
<tr>
<td>CCA [10]</td>
<td>0.902</td>
<td>0.714</td>
<td>0.956</td>
<td>0.532</td>
<td>0.980</td>
<td>0.578</td>
<td>0.902</td>
</tr>
<tr>
<td>LPCCA((k'=15)) [36]</td>
<td>0.954</td>
<td>0.874</td>
<td>0.970</td>
<td>0.128</td>
<td>0.558</td>
<td>0.200</td>
<td>0.960</td>
</tr>
<tr>
<td>LPCCA((k'=30)) [36]</td>
<td>0.954</td>
<td>0.874</td>
<td>0.842</td>
<td>0.072</td>
<td>0.942</td>
<td>0.392</td>
<td>0.960</td>
</tr>
<tr>
<td>MVAE [8]</td>
<td>0.905</td>
<td>0.705</td>
<td>0.974</td>
<td>0.410</td>
<td>0.980</td>
<td>0.598</td>
<td>0.914</td>
</tr>
<tr>
<td>MVAE (with label) [8]</td>
<td>0.908</td>
<td>0.724</td>
<td>0.976</td>
<td>0.432</td>
<td>0.982</td>
<td>0.610</td>
<td>0.920</td>
</tr>
<tr>
<td>SMVAE-UP (the proposed method)</td>
<td>0.964</td>
<td>0.872</td>
<td>0.982</td>
<td>0.758</td>
<td>0.980</td>
<td>0.636</td>
<td>0.970</td>
</tr>
</tbody>
</table>

Fig. 4: Visualization of latent features via t-SNE [37]. Red and blue points show latent features of classes 0 and 1, respectively. The more distant the center of red points and that of blue points are, the higher the discriminative power of the latent features are. Because we denote the distance between the centers of red and blue points by \(d\), we can understand that the discriminative power is high if \(d\) is large.

Table 3: Comparison of classifiers. Columns (a) and (b) denote F-measures for classes 0 and 1, respectively.

<table>
<thead>
<tr>
<th></th>
<th>User 1</th>
<th>User 2</th>
<th>User 3</th>
<th>User 4</th>
<th>User 5</th>
<th>User 6</th>
<th>Ave.</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a)</td>
<td>(b)</td>
<td>(a)</td>
<td>(b)</td>
<td>(a)</td>
<td>(b)</td>
<td>(a)</td>
<td>(b)</td>
</tr>
<tr>
<td>SVM [40]</td>
<td>0.724</td>
<td>0.440</td>
<td>0.934</td>
<td>0.248</td>
<td>0.950</td>
<td>0.084</td>
<td>0.764</td>
</tr>
<tr>
<td>KNN [41]</td>
<td>0.846</td>
<td>0.228</td>
<td>0.962</td>
<td>0.106</td>
<td>0.980</td>
<td>0.001</td>
<td>0.818</td>
</tr>
<tr>
<td>MLP [42]</td>
<td>0.850</td>
<td>0.268</td>
<td>0.964</td>
<td>0.130</td>
<td>0.980</td>
<td>0.002</td>
<td>0.838</td>
</tr>
<tr>
<td>GCN (the proposed method)</td>
<td>0.964</td>
<td>0.872</td>
<td>0.982</td>
<td>0.758</td>
<td>0.980</td>
<td>0.636</td>
<td>0.970</td>
</tr>
</tbody>
</table>

of classes to two and used Euclidean distance.

**MLP (Multi Layer Perceptron) [42]:** We set the number of classes to two. We used ReLU as the activation function and Adam as the optimizer. We determined the number of units in the hidden layer and the batch size via grid search.

Table 3 shows the results. Note that SVM, \(k\)-NN and MLP, unlike our method, do not use a network structure based on
metadata. This table shows the comparatively greater effectiveness of our GCN, which does use the network structure. Through these experiments, we confirmed the effectiveness of our proposed approach, i.e., involving preference-aware multimodal feature extraction based on SMV-AE-UP and then a recommendation based on GCN.

5. Conclusions and Future Work

This paper presented a method for recommending preferred posts to users from a large number of posts. Specifically, we derived SMV-AE-UP, which can extract relationships among content features including consideration of class information that represents whether a user prefers each post. Thus, we were able to obtain preference-aware multimodal features. Furthermore, for each target user, we constructed a network enabling comparison between a user and posts in the same feature space using the preference-aware multimodal features together with metadata of Reblog and users’ following relations. By applying GCN to the constructed network for each target user, it was possible to improve post recommendations that better matched each user’s post preferences.

In this study, we developed two independent methods: SMV-AE-UP for feature extraction and a GCN for recommendation. This approach provides the advantage of applying the preference-aware multimodal features obtained by SMV-AE-UP to many tasks other than recommendations (e.g., classification, retrieval, and visualization). It would be interesting to implement such applications in the future. If we focus only on recommendation applications, an end-to-end neural network including both the feature extraction and recommendation stages would be useful for improving performance. Our future work will include the development of such an end-to-end framework. Furthermore, for the experiments, we need more users from the perspective of the usability for the real-world application. However, there is the limitation because of training time. Specifically, in the experiment, we used the computer with Intel Core i7-7800X, NVIDIA Quadro GV100 and 32GB RAM, and the total time for training SMV-AE was 33.5 hours for all users. Furthermore, we used the computer with Intel Core i7 7700, NVIDIA Geforce 2080ti and 32GB RAM, and the total time for training GCN was 6.88 hours for all users. In the future, we will adopt a fast training method such as Kronecker-factorized approximate curvature [44], then will perform the verification using more users.

Acknowledgment

We thank Deborah Soule, DBA, from Edanz Group (www.edanzediting.com/ac) for critically reviewing and editing a draft of this manuscript. This work was partly supported by the MIC/SCOPE #181601001.

References

27) S. Almutarneh, P. Gamallo, and F. J. R. Pena, “CITUS-COLE at semeval-2019 task 5: Combining linguistic features to identify hate speech against immigrants and women on multilingual tweets,” in Proc. Workshop on Semantic Evaluation,