Weakly Supervised Attention for Hashtag Recommendation using Graph Data

Amin Javari
University of Illinois at Urbana-Champaign
javari2@illinois.edu

Zhankui He
University of California San Diego
zhh004@eng.ucsd.edu

Zijie Huang
University of California Los Angeles
zijiehuang@cs.ucla.edu

Raj Jeetu
University of Illinois at Urbana-Champaign
jraj2@illinois.edu

Kevin Chen-Chuan Chang
University of Illinois at Urbana-Champaign
kcchang@illinois.edu

ABSTRACT

Personalized hashtag recommendation for users could substantially promote user engagement in microblogging websites; users can discover microblogs aligned with their interests. However, user profiling on microblogging websites is challenging because most users tend not to generate content. Our core idea is to build a graph-based profile of users and incorporate it into hashtag recommendation. Indeed, user's followee/follower links implicitly indicate their interests. Considering that microblogging networks are scale-free networks, to maintain the efficiency and effectiveness of the model, rather than analyzing the entire network, we model users based on their links towards hub nodes. That is, hashtags and hub nodes are projected into a shared latent space. To predict the relevance of a user to a hashtag, a projection of the user is built by aggregating the embeddings of her hub neighbors guided by an attention model and then compared with the hashtag. Classically, attention models can be trained in an end to end manner. However, due to the high complexity of our problem, we propose a novel weak supervision model for the attention component, which significantly improves the effectiveness of the model. We performed extensive experiments on two datasets collected from Twitter and Weibo, and the results confirm that our method substantially outperforms the baselines.

CCS CONCEPTS

• **Information systems** → **World Wide Web**; *Information retrieval*; Information systems applications.

KEYWORDS

Hashtag recommendation; Attention mechanism; Scale-free graph

ACM Reference Format:
Amin Javari, Zhankui He, Zijie Huang, Raj Jeetu, and Kevin Chen-Chuan Chang. 2020. Weakly Supervised Attention for Hashtag Recommendation using Graph Data. In *Proceedings of The Web Conference 2020 (WWW '20), April 20–24, 2020, Taipei, Taiwan.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3366423.3380182

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '20, April 20–24, 2020, Taipei, Taiwan
© 2020 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-7023-3/20/04.
https://doi.org/10.1145/3366423.3380182

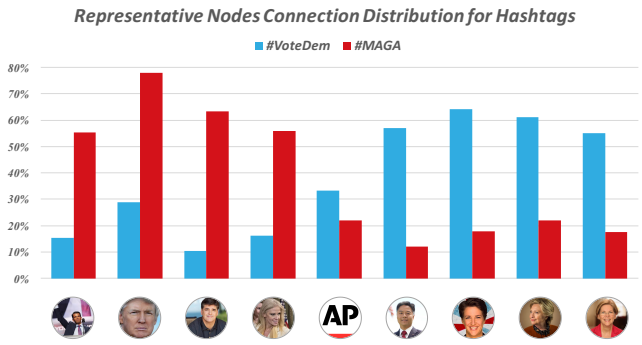


Figure 1: The percentage of links toward 10 representative nodes in the US politics for two groups: users who tweeted hashtag #VoteDem and users who tweeted #MAGA.

1 INTRODUCTION

With an overwhelming amount of *microblogs* spreading in our digital universe through websites like Twitter and Weibo, to help target audience interests, *hashtags* have proven remarkably effective for labeling microblogs. While the usefulness of hashtags has been exploited for many applications (e.g., [12]), a huge and dynamic pool of hashtags exists in microblogging websites which could undermine the effectiveness of the notion of hashtags. As a result, it has been shown that only a small portion of microblogs are annotated with hashtags [23]. Therefore, the task of Hashtag Recommendation for Microblogs, has become indispensable for microblogging websites.

The problem can also be investigated from another perspective: for a user who wants to follow hashtags to find microblogs aligned with her interests, discovering relevant trending hashtags could become a challenge. Although successfully addressing the problem could substantially promote user satisfaction and engagement, it has not received much attention so far in academia. A possible approach could be building content based models, i.e., given the microblogs of a user, the content of the microblogs can be analyzed to identify her interests [25]. Or we could employ classic Collaborative Filtering (CF) models to make recommendations by analyzing historical user-hashtags interactions [1]. However, only a small portion of users

RIGHTSLINK

tend to generate abundant content data and interact with hashtags which strongly affects the applicability of such approaches [28].

Problem: Given a user on a microblogging website like Twitter, how can we recommend her hashtags, in particular when sufficient content/interaction data is not present?

We propose that the problem can be approached by relying on graph data. In fact, it has been shown that 1) a portion of links can be regarded as interest-based links [5, 26], 2) and also link data is more abundant than content data on such websites since most of the user tend to be 'listeners' [28].

However, *how do we exploit users' social linkage in an effective way to address the problem?* A classic approach could be profiling a set of users in the network based on their content/interaction data and then propagating their embeddings in the network through recent methods like graph convolutional networks [38] or graph attention networks [35] to build the profiles of other users. Nevertheless, such methods do not fit our recommendation problem due to two major reasons:

Efficiency: It is computationally expensive to apply them in multiple iterations to a huge network like Twitter. In fact, even crawling a network with millions of nodes and billions of edges has burdensome computations. Also, microblogging websites are highly dynamic and a large number of nodes/edges emerge on a daily basis which means that the profiles of users need to be updated frequently.

Effectiveness: Such models build a fixed profile for each user. In a microblogging website, a user may have diverse and independent interests such as politics and sports. To determine whether a hashtag related to politics is relevant to the user, it is more effective to build a profile that focuses on her political interests, i.e. , involving unrelated interests can potentially add noise in the prediction process. In fact, this is the core idea behind most of the CF based models, including item-based CF [20].

Towards building a model that addresses the mentioned drawbacks of a classic graph based encoding technique, we aim to take advantage of the structural characteristics of the microblogging networks. Microblogging networks are scale-free graphs [2, 4, 28]. There are a set of hub nodes that receive a large portion of the other users links [4]. Also, it has been shown that hubness/popularity of a user on Twitter indicates that she represents a topic of interest [28]. That is, the links towards hub nodes effectively capture user's interests. As such, our insight to build an efficient model is to focus on user's links towards hub nodes rather than analyzing the entire network. Fig. 1 illustrates the intuition behind the informativeness of hub nodes in determining what hashtags a user might be interested in. It depicts the distributions of the links that two sets of users have towards a small set of hub nodes in the US politics in which followers of a republican leader are more likely to use hashtag #MAGA while the followers of a democrat leader tend to use hashtag #VoteDem.

Based on this insight, our key idea is that in a microblogging website, the embeddings of nodes can be derived by aggregating the embeddings of the hub nodes they are connected to. That is, rather than finding the embeddings of entire nodes in the network, we can obtain the embeddings of hub nodes and build a function to transitively embed other nodes. We use this general idea as the basis of our hashtag recommendation model, i.e., we pick a set of hub users denoted as **representative nodes** and project those nodes and hashtags into a shared latent space. Given a user and a hashtag, we construct an embedding of the user based on her representative followees and then feed it to a scoring model along with the embedding of the target hashtag to generate the relevance score.

The proposed architecture meets the efficiency challenge. It relies on profiling a very small portion of the nodes and transitively derives the profiles of other users. Also, it deals with the high dynamics of the network. Unlike ordinary users, profiles of representative-nodes have low dynamics, and the dynamics of the emergence of such nodes are quite slow. The embeddings of users are defined as *functions* of representative-nodes profiles. Hence, we can capture the frequent changes in user's profiles by simply changing the function's inputs. Moreover, the model is inductive as it can make recommendations for unseen/new users.

More importantly, it enables us to use hashtag-aware embeddings of users in the recommendation process, hence addresses the effectiveness challenge. To embed a user based on her representative followees, we propose a model based on attention mechanism which has been successfully employed in different problems [9, 20]. Given a hashtag, our proposed attention model embeds the target user by performing a weighted aggregation of her followees' embeddings where the weights reflect the relatedness of a hashtag to a representative node.

However, employing the idea of attention mechanism in the proposed model is challenging. Classically, an attention model can be learned implicitly, i.e., in an end to end manner based on the final objective of the model. However, due to sparsity issue caused by a large number of free parameters in our model (a huge pool of representative nodes and hashtags), the attention model learned by implicit training generates low quality attention maps. To tackle this issue, as a general approach, explicit supervision can be injected into attention models using labels specifically acquired for the attention model [9]. However, the challenge we encounter is that the labels for relevance of representative nodes to hashtags do not exist. In response, we propose the idea of weak supervision, i.e., using statistical methods, we first generate weak labels for the relevance of hashtags to representative nodes and then use the labels as explicit supervision for the attention component.

Lastly, to show the flexibility of the proposed graph based model, we further develop the model to accommodate content (text) data. We conducted comprehensive experiments on two datasets obtained from Twitter and Weibo and compared the performance of the model with multiple baseline recommendation models. Our results show that the proposed model substantially outperforms the baselines in terms of NDCG and HR. Also, the results confirm the considerable effectiveness of the proposed supervised attention based technique over the unsupervised version of the model. The major contributions of the paper can be summarized in two perspectives:

- We introduce the problem of trending hashtag recommendation based on graph data.
- We introduce the novel idea of node embedding based on representative/hub nodes which is effective for large scale dynamic graphs like Twitter or Weibo.

RIGHTSLINK

- We propose the notion of weakly supervised attention in a graph setting to effectively encode users based on their neighbors in the context of a hashtag or topic.

2 PROBLEM DEFINITION AND PRELIMINARIES

We first formalize the problem and then shortly recapitulate the widely used Generalized Matrix Factroization model which is the basis component of our proposed model.

2.1 Problem Formulation

Let $G(V, E)$ denote the user's directed connections in a Microblogging website where V is the set of nodes/users, and E represents the set of the links. And let U denote the target set of users with the size M where $U \subseteq V$, H the set of hashtags with size N, and T_u the set of the tweets of user u. We define the user-hashtag interaction matrix as $\mathbf{Y} \in \mathbb{R}^{M \times N}$ where $y^{uh} = 1$ if an interaction is observed between u and h, e.g., u has used h in one of his tweets T_u, and $y^{uh} = 0$ otherwise. Note that a value 0 does not necessarily mean u is not related to h, it can be that the user is not aware of the hashtag or the user is not active generally. Also, it should be mentioned, unlike a traditional recommendation problem, a large proportion of users within U do not have any interactions with H. The recommendation problem for $u \in U$ is formulated as the problem of estimating the scores of unobserved entries in the corresponding row of $\mathbf{Y}$, which are used for ranking the hashtags for u.

2.2 Generalized Matrix Factorization

Matrix Factorization can be considered as the most popular model for recommendation and has been investigated extensively in literature [7]. Matrix Factorization based models define the relevance score of user u and item i as the inner product of their latent vectors $\hat{y}_{ui} = p_u^T q_i$ where p_u and q_i represents the latent vectors of u and i respectively. Generalized Matrix Factorization is a recent model that uses a multi-layer neural-network structure to mimic classic MF [21]. In GMF, the input layer consists of two feature vectors v_u and v_i that describe user u and item i, respectively. Above the input layer is the embedding layer; it is a fully connected layer that projects the sparse representation of the target user and item to dense vectors. The obtained user/item embedding can be seen as the latent vector for user/item. The user embedding and item embedding are then fed into a neural architecture to estimate the relevance scores.

3 PROPOSED MODEL: PHAN

In this section, we describe the structure of the proposed model denoted as Personalized HAshtag recommendation based on Netwrok data (PHAN). In the following, first, the notion of representative nodes is illustrated as the core idea of the model; next, the model is introduced in a top-down manner: we start with the prediction model, and then the encoding techniques are described; finally, we explain how the model can be altered to capture both text and graph data. Fig. 2 represents the structure of the model based on both graph and text data.

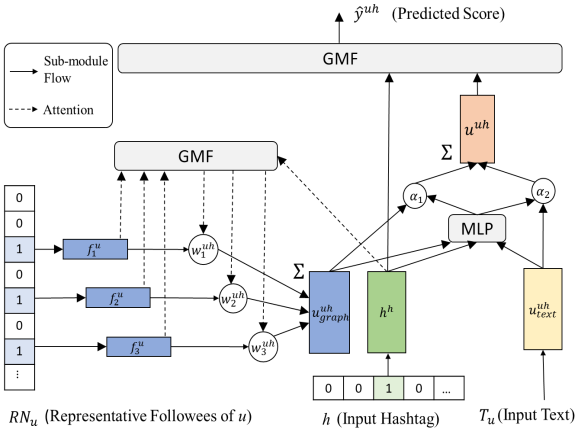


Figure 2: Architecture of PHAN based on graph and text data. The embeddings of the target hashtag, h^h, and the embeddings of the u's representative follows, RN_u, are obtained from lookup tables. The graph based embedding of u with respect to h, u^{uh}_{graph}, is the weighted aggregation of embeddings of u's representative followees where the weight of the embedding of node j, w_j^{uh}, is determined by feeding f_j^u and h_h to the GMF function. u's text-based profile, u^{uh}_{text}, is built by an off the shelf embedding model. u^{uh}_{text} and u^{uh}_{graph} are combined in a weighted way to build a unified profile of the user, u^{uh}, where the weights are determined using an MLP. Finally, $\hat{y}^{uh}$ is generated by feeding h^h and u^{uh} to the GMF.

3.1 Representative Nodes as User Features

To predict the relevance of user u to hashtag h, we rely on GMF [21]. The feature vector of users and items in GMF can be customized to support a wide range of modeling of users and items, such as content-data, user-item interaction, or context data. In our problem setting, content-based or interaction-based features are not applicable because most of the users are 'listeners' ($T_u = \emptyset$).

We propose to exploit user's links on graph G to profile users based on the fact user's links implicitly indicate their interests, and also graph data is more abundant than content data [5]. However, it is of great importance to build the model in a way that it is *efficient* and *effective*. As mentioned, a general graph embedding technique does not meet this criterion. Hence, we aim to address the problem by introducing the notion of *representative nodes*.

3.1.1 Why Representative Nodes? There are two main reasons for the creation of a link from an initiator node to a receiver node in a microblogging website. 1) The receiver node is associated with a topic interesting for the initiator node. 2) There exist a friendship/social tie between the nodes [5, 28]. A user can receive only a limited number of social tie based connections. This implies that when a node becomes popular in a microblogging network, it boldly represents a topic of interest. In fact, there exist some popular representative node for each topic of interest, and if a user is interested in a topic, she very likely follows some of the corresponding representative nodes [2].

RIGHTSLINK

Based on this generative process of links in microblogging websites, we observe that Microblogging networks are scale-free networks, i.e, they contain some hub nodes and a large portion of links are toward hub nodes [2]. Also, those links are noiseless and purely indicate user's interests (unlike the links among ordinary users which are likely to represent friendship ties) [5]. In accordance with the power-law distribution, the number of hub nodes is extremely smaller than the number of nodes in the network [4]. Also, representative nodes have slow dynamics. They normally appear or change in the network gradually over time. Moreover, the interest topics associated with a representative node are mostly fixed except for rare cases, e.g., when a popular athlete becomes a politician. Lastly, it is practically straightforward to identify topics represented by representative nodes and profile them as they are associated with ample data (both content and graph data).

These characteristics of representative nodes make them a noiseless, low dynamic, and dense data source in general to profile users and in particular, to address our underlying problem. As such, we aim to use the list of representative nodes followed by a user as the feature set describing the user. We emphasize the idea of user profiling based on representative nodes heavily depends on the fact that the input network is a scale-free network.

3.1.2 Selecting Representative Nodes. As the initial step of the model, the set of representative nodes is selected. Given the target set of users U and a graph $G = (V, E)$, we define representative nodes $RN = \{r_1, r_2, ..., r_n\}$ as the set of nodes in V with at least L followers in U where n is the number of representative nodes. Accordingly, the set of representative followees of the user u is represented as $RN_u \in RN$. It should be noted, L is a parameter of the proposed model to be tuned according to desired performance and efficiency. In general, the value of L is decided by the size of U. We further discuss this in the experiments section.

3.2 Predicting Relevance Score

Following the structure of GMF, given the embedding vector of the target user $\mathbf{u}^{uh}$ and the target hashtag $\mathbf{h}^h$, the relevance between u and h is computed as follows:

$$\hat{y}^{uh} = \sigma(\mathbf{W}^{\mathbf{T}}(\mathbf{h}^{\mathbf{h}} \odot \mathbf{u}^{\mathbf{uh}}) + b), \quad (1)$$

where σ is Sigmoid function $\sigma(x) = \frac{1}{1+e^{-x}}$, $\mathbf{W}$ and b_{fc} are the weight and bias of GMF, $\odot$ denotes element-wise multiplication. Note that, the GMF function can be replaced with other prediction models such as multi-layer perceptron (MLP) as in [21]. In the next subsections, the embedding components of the model are introduced.

Hashtag Embedding: In our problem settings, hashtags are not associated with any extra data but identities. Hence, we represent hashtags as one hot vectors of identities and learn their representations by building a lookup table which can be learned in our training process. Note that additional data can also be incorporated to embed hashtags.

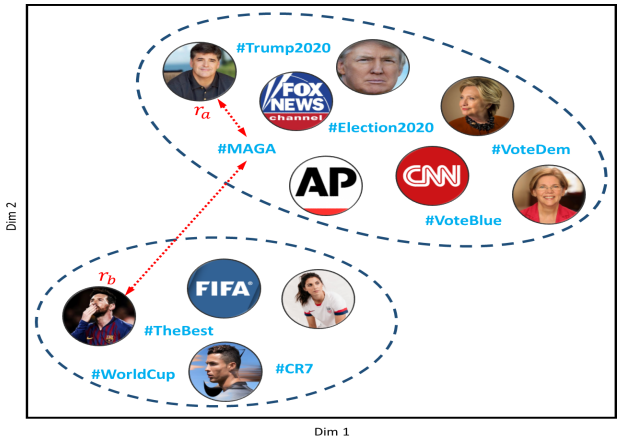


Figure 3: A toy example representing the hashtags and representative nodes related to politics and soccer in a shared latent space

3.3 User Embedding based on weakly supervised attention

We propose a weakly supervised attention based graph embedding technique to embed user u based on RN_u to determine its relevance to hashtag h. In the following, we first introduce a basic approach to encode the followee list of users to better motivate the proposed model and then illustrate the architecture of the model.

Design 1: Fixed embedding. A basic approach to encode a user based on her following list is to create a one-hot encoded vector where each element corresponds to a representative node and is marked 1 if the user follows the node and 0 otherwise. The resulting sparse vector can be projected to a fully connected dense embedding vector.However, this trivial approach is oblivious to the target hashtag information and creates a single fixed representation for the user. A user might have a set of independent interests, e.g., sports, politics and music. When predicting the relevance of a user to a hashtag, it is more accurate to focus on the part of the user's interests that is related to the target hashtag.

Design 2: Embedding based on classic attention. Towards building a hashtag dependent representation of a user, we suggest that hashtags and representative nodes can be projected into a shared latent space. Roughly speaking, we assume if user $u \in U$ follows representative node r and uses hashtag f in one of his tweets, it indicates proximity between r and f in the embedding space. Having the embeddings of representative nodes and hashtags, the hashtag dependent embedding of the target user can be obtained by aggregating the embeddings of her representative followees in a weighted way where the weights are determined based on the relevance of the representative nodes to the target hashtag.

We further explain the idea behind our target aware embedding model using the toy example depicted in Fig. 3. The figure shows the representative nodes and hashtags of two topics: politics and soccer in a shared latent space. Let's assume the user u follows some representative nodes related to politics, including the representative node r_a and some representative nodes related to soccer,

RIGHTSLINK

including the representative node r_b. Suppose the task is to determine the relevance of u to hashtag #MAGA. We suggest that to embed u with respect to #MAGA, the attention model should attend more to the representative node r_a than r_b because r_a has higher proximity to hashtag #MAGA in the embedding space. In this way, the model focuses on the politic related profile of the user to make the prediction and filters out the unrelated parts.

More formally, the attention-based model can be described as follows: let's assume we have a lookup table for the embeddings of representative nodes which can be learned in an end-to-end manner based on the final output of the model. We can look up the embedding table and get the embeddings of the representative nodes followed by the user u denoted by $\mathbf{F}^u \in \mathbb{R}^{s \times n}$ and hashtag embedding denoted by $\mathbf{h}^h \in R^{s \times 1}$, where s is the size of the embedding vector, and n is the number of representative nodes followed by u. We define the user embedding as:

$$\mathbf{u}^{uh}_{graph} = n^{-\beta} \sum_{j \leq n} \hat{w}^{uh}_j \cdot \mathbf{f^u_j}, \tag{2}$$

where β is a normalization parameter ranging from 0 to 1 and $\mathbf{f^u_j}$ represents the embedding of the j-th representative node in $\mathbf{F}^u$, i.e., $\mathbf{F}^u =< \mathbf{f^u_1}, ..., \mathbf{f^u_n} >$. Also $\hat{\mathbf{W}}^{uh} =< \hat{w}^{uh}_1, ..., \hat{w}^{uh}_n > \in \mathbb{R}^{n \times 1}$ is the set of attention weights for user u where $\hat{w}^{uh}_j$ denotes the attention weight of the j-th representative node with respect to hashtag h.

We employ GMF-based prediction model to compute the attention weight for each representative node given the target hashtag. As such, given a representative node and the target hashtag, the dependency between them can be computed as follows:

$$\hat{w}^{uh}_j = \sigma(\mathbf{W^T}(\mathbf{h^h} \odot \mathbf{f^u_j}) + b). \tag{3}$$

It should be noted that the GMF model we use in our attention model shares the same parameters as the GMF model for relevance prediction in because a target user and representative nodes share the same embedding space, hence parameter sharing can be used for the GMF models.

Having this structure, the attention model, and the embeddings of representative nodes can be trained in an end to end manner based on the final objective.

However, the success of the proposed graph embedding technique and, consequently, the accuracy of the prediction model depends on the effectiveness of the attention model. The more accurately attention determines the relatedness of representative nodes to a given hashtag, the more accurate encoding of the target user can be obtained. However, according to our experiments, the accuracy of the second design is not superior to the first one. We conjecture the reason for this is that the proposed attention model does not accurately determine the relatedness weights.

Indeed, attention-based models have been used for various problems such as image captioning and Visual Question Answering (VQA) [10, 31], e.g., in VQA, attention is used to selectively target related areas of an image to a given question. Classically, the attention mechanism can be trained in an end to end manner based on the final objective of the model. Generally, this form of training is effective for models/problems with low complexity. However, the accuracy of the attention model that is implicitly trained is not ensured, especially for models with high complexity. For example, in an inherently complex problem like VQA [31], attention models trained implicitly do not focus on the same regions as humans do, which results in incorrect answers.

Why classic attention does not generate accurate attention maps? The proposed attention model can be considered as a sub-component of the main prediction model. When training the model in an end-to-end manner, the parameters of the attention model are learned based on the final output. However, this way of training is quite vulnerable to the vanishing gradient problem [29]. In general, the vanishing gradient problem occurs during back-propagation in a deep neural network when the magnitude of gradients diminishes as we go down the layers. This causes convergence to occur extremely slowly as the updates are highly diminished. Also, overfitting is another issue associated with attention models. Due to the large size of the set of hub-nodes and hashtags, we encounter overfitting/sparsity in classic-attention. Some complex attention-models used in VQA had also faced the problem [31]. In our model, if the embeddings of users are obtained by equally attending over representative nodes (fixed embedding, without target aware attention), the model would have less complexity with a lower degree of freedom. However, applying the attention component to the model adds up to its complexity. Hence, training the model would become more prone to overfitting.

Design 3: Embedding based on Weakly Supervised Attention. In order to tackle this challenge, some recent works have tried to add supervision to the attention component of their models. For example, in a recent VQA model, when human-labeled training data is used for supervising the attention component, not only the accuracy of question answering increased, but also the error of the attention model is reduced [11, 31]. Inspired by the success of the general idea of supervised attention, we aim to add attention supervision to our model. However, we face a major challenge: the labels of relevance for a representative node and a hashtag do not exist.

In response, we introduce the idea of *weak supervision for attention mechanism*. We propose that *weak labels* for the relevance of nodes and hashtags based on training data and by relying on statistical models. We introduce two types of supervision based on this idea and inject them into the final loss function of the model. This allows us to jointly supervise the attention and relevance estimation models in an explicit way.

Co-occurrence based Supervision: The main idea behind our first weak supervision is that if a user uses hashtag h in her microblog and also follows r_i, it reflects a relevance between h and r_i. That is, we regard the attention component as a link prediction model: if a user is associated with hashtag h, how likely she follows the target representative node? Accordingly, the desired output of the prediction model w^{uh}_j in $\mathbf{W}^{uh} =< \hat{w}^{uh}_1, ..., \hat{w}^{uh}_n > \in \mathbb{R}^{n \times 1}$ is defined as $P(h|j)$, whose Maximum Likelihood Estimate is as follows:

$$w^{uh}_j = P(j|h) = \frac{c(h,j)}{c(h)}, \tag{4}$$

where $c(h, j)$ is the co-occurrence count of hashtag h and representative node j in the training data and $c(h)$ is the frequency of hashtag h in the training data.

RIGHTSLINK

Informativeness based Supervision: This form of weak supervision is inspired by the concept of TF-IDF in information retrieval. We suggest that the relevance of representative node r_i and hashtag h is decided by two factors: 1) the number of users associated with hashtag h who also follow r_i. 2) the number of users who follow r_i and are associated with hashtags other than h. Through this idea, we penalize those popular nodes that are associated with a large number of hashtags. Following this idea, w_j^{uh} is defined as:

$$w_j^{uh} = \frac{c(h,j)}{c(h)} * \log \frac{c(j)}{c(h',j)}, \tag{5}$$

where $c(h',j)$ is the number of times node j has occurred in the dataset but in absence of hashtag h while $c(j)$ denotes the number of times node j occurs.

3.4 Loss function

As mentioned, we inject supervision to the attention model through our loss function. Hence, the loss function of our proposed model consists of two parts: the main loss function and the supervised-attention based loss function.

Main Loss Function: The main loss function is defined based on the difference between the estimated relevancy score $\hat{y^{uh}}$ and ground-truth label information y^{uh}. In our task, y^{uh} is 1 or 0 according to whether the user u has interacted with hashtag h. We define the loss function as the Binary Cross-Entropy Loss Function.

Supervised-Attention based Loss Function: We introduced Attention supervision to guide our GMF model to better estimate the relevance between the target representative node and hashtag, where the supervision information w_j^{uh} in $\mathbf{W}^{uh} =< \hat{w}_1^{uh}, ..., \hat{w}_n^{uh} > \in \mathbb{R}^{n\times 1}$ is represented as $P(j|h)$, which means given a hashtag h, how relevant the followee j is. We define the loss function between w_j^{uh} and $\hat{w}_j^{uh}$ as the Binary Cross-Entropy Loss Function as well.

Joint Loss Function: We define the joint loss function as:

$$\begin{aligned} Loss =& L_{main}(\hat{y}^{uh}, y^{uh}) + \lambda L_{attention}(\hat{\mathbf{W}}^{uh}, \mathbf{W}^{uh}) \\ =& -(\hat{y}^{uh} \log y^{uh} + (1-\hat{y}^{uh}) \log(1-y^{uh})) \\ & - \frac{\lambda}{n} \sum_{j \le n} \hat{w}_j^{uh} \log w_j^{uh} + (1-\hat{w}_j^{uh}) \log(1-w_j^{uh}), \end{aligned} \tag{6}$$

where λ is a trade-off setting to show how important the role attention supervision plays in this model. When $\lambda = 0$, the model can be regarded as a classic attention model.

How joint loss function improves the effectiveness of the attention model? 1) By providing explicit supervision for a shallow sub-component of the model, our approach helps to mitigate the vanishing gradient problem. 2) Further, attention supervision acts as a regularizer in the loss function. Hence, it can help prevent overfitting in training the model. Indeed, it adds more constraints to the parameters of the attention model. 3) To build a profile of a user, it is more reliable/robust to attend to representative nodes with more reliable embeddings. The supervised attention based loss function automatically lowers the contribution of less reliable representative nodes. Statistically, the embedding obtained for a representative node is more reliable if it has been obtained based on more data, i.e., if it has a larger number of co-occurrences with the set of hashtags. According to equations 4 and 5, in average, less reliable representative nodes will have less relevance to the hashtags. As such, they have a smaller contribution in building the representation of a given user.

3.5 Fusing content and graph data

Although most of users are not associated with content data, for active users, abundant content data is available, which can be used to profile them. While we focus on exploiting graph data– as it is orthogonal to text/content data– the knowledge extracted from text data can be fused with the profile obtained from the target user's connections to build a unified recommendation model for users. Furthermore, fusing text and graph data enables us to use the model as a personalized hashtag recommendation model for *microblogs*. That is, the knowledge extracted from a user's microblog can be combined with her graph-based profile to make personalized recommendations for the target microblogs. In fact, having the profile of a user can aid us to better understand the content of her microblogs and resolve their inherent ambiguities, which consequently improves the quality of recommendations.

In both of the scenarios, User-Hashtag recommendation (UHR) and Microblog-Hashtag recommendation (MHR), the recommendation problem can be defined as finding the relevance score between a hashtag h and a pair of text-followee list $< T_u, RN_u >$. Note that if we use the model as a personalized hashtag recommendation model for microblogs, text data would involve a single microblog, i.e., $T_u = \{t\}$ where t is the target microblog.

Text/content and graph data are two different data sources in nature: the former is sequential while the latter is associative (non-sequential). Considering their inherent differences, we propose to encode text and followee data separately and then fuse them to build an embedding of the input data, which can be fed to the prediction model described in Eq. 1 to solve the target relevance prediction problem. Various techniques have been introduced in the literature to combine embedding vectors [42]. Based on our experiments, we opt to combine the two output vectors using weighed element-wise addition. We chose to infer the weights using the hashtag information based on the intuition that the importance of the two channels differs depending on the hashtag.

In more details, given hashtag h the unified embedding vector denoted by u^{uh} is obtained as follows:

$$\mathbf{u}^{uh} = \alpha_1 \mathbf{u}_{text}^{uh} + \alpha_2 \mathbf{u}_{graph}^{uh}, \tag{7}$$

where α_1 and α_2 are hashtag-dependent weights and $\mathbf{u}_{text}^{uh}$, $\mathbf{u}_{graphl}^{uh}$ are text and followee list embeddings respectively. To predict the weight scalars α_1, α_2 with regard to the input hashtag, we employ a multi-layer perceptron (MLP) as the weight prediction model with hashtag, text and followee list embeddings as input.

In previous subsections, we described how u_{graph}^{uh} can be obtained. Now the question becomes how to encode the text data? Our fusion architecture allows us to adopt the existing text embedding techniques. We apply two state-of-the-art microblog embedding techniques to exploit textual information. For the task of Microblog-Hashtag recommendation(MHR) , we apply the classic Long-Short-Term Memory (LSTM) [22] model to encode the single input microblog, which is a widely used technique for embedding short text

RIGHTSLINK

data. Meanwhile, for User-Hashtag recommendation(UHR), the end-to-end Memory model [24] is used due to its excellent performance in handling multiple microblogs.

4 EXPERIMENT

We conducted experiments to verify the effectiveness of our proposed model. We focus on answering three key questions:

- **RQ1:** How does our model perform compared to the baseline methods on UHR and MHR tasks?
- **RQ2:** How effective is the proposed attention model?
- **RQ3:** Do users have sufficient number of links towards representative nodes?

4.1 Experimental Setup

Dataset: We experimented with two real-world datasets: Twitter Dataset and Weibo Dataset. Related statistics are described in Tab 1 where **RN** denotes Representative Node.

Twitter Dataset: Based on a Kaggle dataset[1] that contains Twitter user's followee list, we use Twitter API to crawl users' public tweets from Jan-2016 to Jun-2016 so as to construct the Twitter Dataset. We selected hashtags whose frequency is more than 30 and representative nodes who are followed at least by 120 users.

Weibo Dataset: We used the public Weibo dataset[2] collected from Sina-Weibo for evaluation. Similar to the Twitter Dataset, we retained hashtags with frequency more than 6 and representative nodes with at least 40 followers.

Evaluation: We use Hit Ratio (HR@K) and Normalized Discounted Cumulative Gain (NDCG@K) [19] as metrics. HR@K is defined as:

$$HR@K = \frac{\#hits@K}{\#tests} \tag{8}$$

where $\#tests$ is the total size of testing data. In Microblog-Hashtag recommendation, it means the number of testing tweets, and in User-Hashtag recommendation, it's the number of testing users. $\#hits@K$ is the number of hashtags which were successfully recalled in the top-K hashtags ranking list. And NDCG@K is defined as:

$$NDCG@K = \frac{DCG@K}{IDCG@K} \tag{9}$$

where $DCG@K = \sum_{i=1}^{K} \frac{2^{rel_i}-1}{log_2(i+1)}$ to accumulate the graded relevancy rel_i of hashtags at position i. $IDCG@K$ is the $DCG@K$ score of perfect ranking list for normalization. In our experiments, we set rel_i as binary relevancy as [19] did.

The dataset was split randomly into two parts with ratio 8:2 as training set and testing set respectively. Validation set was used for tuning the hyper-parameters and later combined with training for final model. As [21] did, for each positive interaction, we sampled 99 hashtags that have no interaction with this user (for User-Hashtag recommendation) or this microblog (for Microblog-Hashtag recommendation), and evaluate the ranking list among these hashtags.

Baselines: This paper is mainly focused on UHR task. As mentioned UHR solely based on graph data has not been investigated in previous works. As such, we compare the model with baselines that perform based on historical user-hashtag interaction data and textual data. The following is a brief description of the baselines were used for the UHR task.

- **Classification based Recommender (CBR):** Analogous to [8], we model the recommendation task as a multi-class classification problem for textual data with word frequencies as textual features.
- **LDA based Recommender (LDAR):** [27] employs a model based on topic modeling over textual data to recommend hashtags.
- **LSTM based Recommender (LSTMR):** In this method, content of microblogs are encoded using LSTM and then obtained encoding is assigned to a hashtag [22].
- **Matrix Factorization (MF):** [1] feeds user-hashtag historical interaction data to a Matrix-Factorization based model to make hashtag recommendations for users.
- **Memory Network based Recommender (MNR):** [24] is a novel end-to-end memory network to model users based on the content of their microblogs. We have adopted the model for UHR task.

It should be noted, **CBR**, **LSTMR** and **LDAR** have been originally proposed for MHR task. We adopt them for the UHR task. In fact, instead of feeding a single microblog, a set of microblogs were fed to the models.

Moreover, to show the effectiveness of the model as a personalized recommendation model for MHR task, we compared it with text based MHR models including **CBR**, **LSTMR** and **LDAR**. Additionally, we used the following state of the art text based models as baselines:

- **CNN-Attention based recommender (CNNAR):** [17] employs a convolutional neural network based architecture equipped with an attention mechanism to effectively analyze the content of microblogs for the hashtag recommendation task.
- **EmTaggeR:** [13] makes recommendations by finding the relevance of the embedding of the target microblog to that of the candidate hashtag. Embeddings of hashtags and microblogs are obtained based on the embeddings of the constituent words.

It should be noted that idea of building personalized MHR models by relying on demographic, personal and location based information has been investigated in previous works [16, 36, 41]. However, such user features are not available in our dataset. Also, the personalization methods introduced in those works are orthogonal to PHAN.

4.2 Performance of proposed model (RQ1 & RQ2)

The overall performance of PHAN and baselines over Twitter and Weibo datasets are displayed in Tab 2. Also, we evaluated four different variations of the PHAN model: 1) Supervised attention based embedding based on co-occurance (**PHAN-SACO**), 2) Supervised Attention based embedding based on Informativeness (**PHAN-SAIN**), 3) Classic attention based embedding (**PHAN-CA**)

[1] https://www.kaggle.com/hwassner/TwitterFriends/home
[2] https://www.aminer.cn/weibo-net-tweet

RIGHTSLINK

Dataset	#Tweet	#Followees	#RN	#Hashtag	#User	Avg(#RN$_{user}$)
Twitter	217965	253818	9309	2873	23169	33.42
Weibo	10521	585475	8426	2017	7023	121.98

Table 1: Statistics of the Datasets.

User-Hashtag recommendation					Microblog-Hashtag recommendation				
	Twitter		Weibo			Twitter		Weibo	
Model	HR@10	NDCG@10	HR@10	NDCG@10	Model	HR@10	NDCG@10	HR@10	NDCG@10
CBR	19.83%	8.43%	17.21%	7.64%	CBR	50.89%	24.97%	39.74%	29.12%
LDAR	18.74%	7.97%	15.96%	6.87%	LDAR	43.79%	22.03%	34.06%	27.43%
LSTMR	23.14%	11.59%	39.47%	27.22%	LSTMR	56.21%	38.31%	52.43%	35.47%
MF	26.30%	14.98%	23.26%	12.29%	EmTaggeR	62.47%	49.69%	57.63%	44.09%
MNR	28.29%	17.00%	30.92%	20.75%	CNNAR	63.64%	49.48%	55.95%	43.71%
PHAN-FIXD	41.21%	24.33%	38.43%	24.78%	PHAN-FIXD	74.65%	55.76%	61.58%	51.40%
PHAN-CA	41.35%	24.25%	38.46%	24.83%	PHAN-CA	73.46%	53.79%	61.70%	51.42%
PHAN-SACO	42.79%	26.20%	40.45%	26.48%	PHAN-SACO	74.64%	55.27%	63.41%	51.54%
PHAN-SAIN	**43.73%**	**26.33%**	**42.06%**	**27.94%**	PHAN-SAIN	**77.57%**	**58.54%**	**65.31%**	**52.94%**

Table 2: Performance of the proposed model (PHAN) and the baseline methods on the Twitter and Weibo datasets

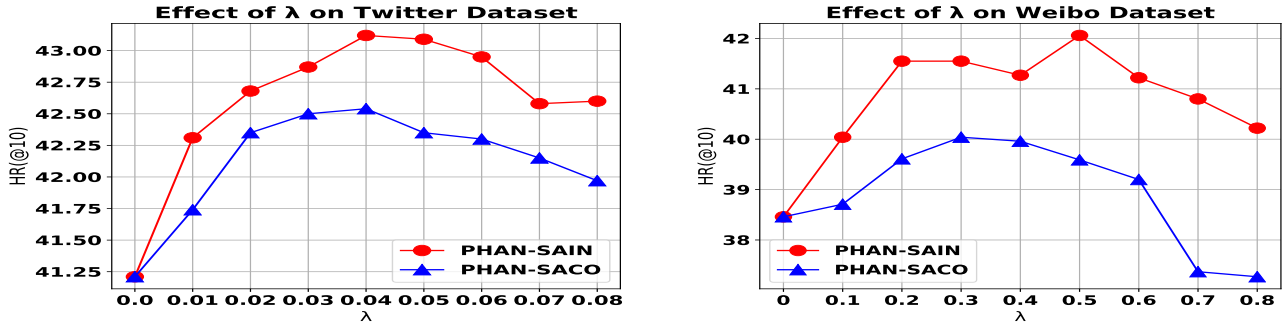


Figure 4: HR as a function of λ on Twitter and Weibo Datasets

and 4) Fixed Embedding (**PHAN-FIXD**). In PHAN-CA, we set $\lambda = 0$ in the joint objective function, i.e., the supervision for attention is not applied. In PHAN-FIXD, following the structure introduced in in [21], we fed the input followee list to the network to get the fixed embedding of the user (hashtag-independent). We used both text and graph data in all four variations.

1) As it can be seen in Tab. 2, PHAN with supervised attention outperforms the other variations. For example, in MHR task, PHAN-SAIN outperforms the PHAN-FIXD model on the Twitter dataset by 2.92% in terms of HR@10; For UHR task, we observed the same behavior whereas on the Weibo dataset PHAN-SAIN outperforms PHAN-FIXD and PHAN-CA by at least 3.5% in terms of HR@10. The results confirm the effectiveness of weakly supervised attention. In fact, while the fixed embedding model has comparable results to the classic attention model, applying the supervision significantly improves the results. Also, PHAN-SAIN provides better results than PHAN-SACO, which means that the form of weak supervision is of great importance.

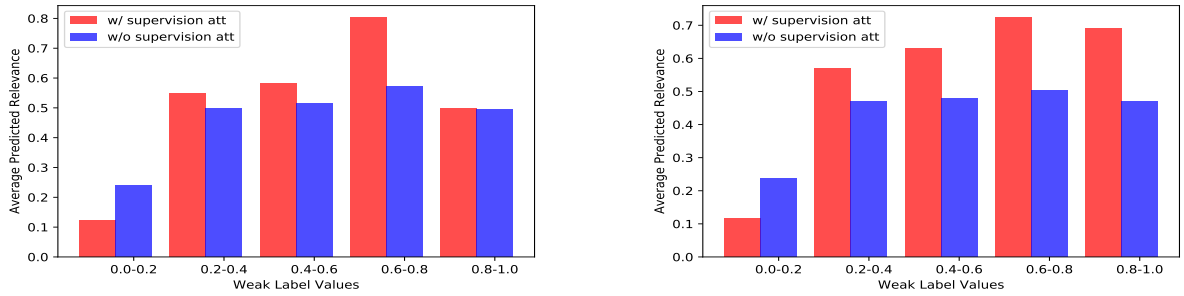
2) The results strongly confirm the effectiveness of incorporating link data in hashtag recommendation for UHR. PHAN outperforms all baselines which consider only text data. For example, PHAN-SAIN outperforms an advanced text-based profiling model like MNR by 15.44% and 11.14% in terms of HR@10 on the Twitter and Weibo datasets respectively. It should be mentioned, the experiments were done only on users who have associated text data. However, a large portion of users do not generate content, hence, text based models are not applicable to them.

3) Also for MHR, we observe that the idea of personalized recommendation based on graph data is quite effective. For example, in MHR task, PHAN-SAIN outperforms CNNAR by 12.94% in terms of HR@10 on the Twitter dataset.

4.3 Effectiveness of supervised attention (RQ2)

In this first section of this experiment, we investigated the model's performance as a function of λ, which determines the importance of attention supervision during training. Figure 4 shows the effect of increasing λ for UHR task for both of our supervision models. We

RIGHTSLINK



(a) Twitter (b) Weibo

Figure 5: Average relevancy value between embeddings of representative nodes and hashtags with and without supervised attention for different bins of weak labels

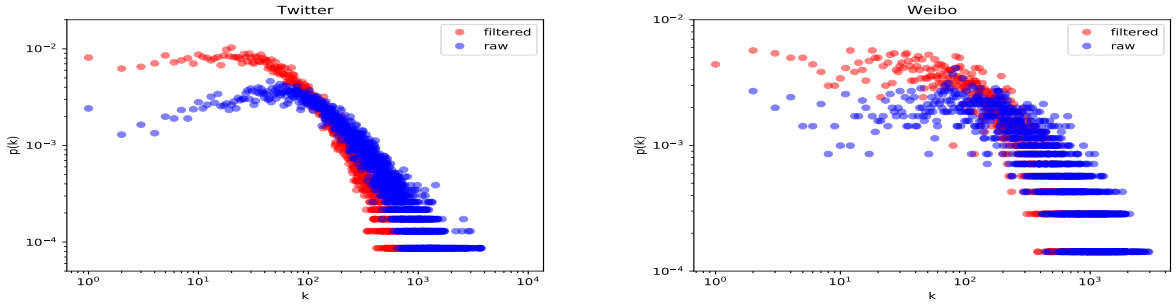


Figure 6: Distribution of the outdegree of users towards representative nodes (filtered) and towards entire set of nodes (raw)

observe two behaviors in these figures. Firstly, the variation trend of $HR@10$ with varying λ is quite stable across the two datasets. Initially, as we increase λ, $HR@10$ increases but later it declines. This is because higher values of λ misled the models' objective function to fit the relevancy between representative nodes and hashtags rather than to fit the relevancy between targeted users with hashtags. Secondly, the optimal λ differs for two datasets.

In the second part of the experiment, we show how supervised attention aids the model to better determine the relevance of representative nodes and hashtags. To this end, we first find the co-occurrence based weak labels for a set of hashtag-representative node pairs and split the obtained values into multiple bins. Next, we find the relevance score of the set of the pairs within each bin by feeding their embeddings to the GMF function. The embeddings are obtained based on two versions of the model: without supervised attention and with co-occurrence based supervision. The figure 5 shows the average relevance score of the embeddings of hashtag-representative nodes pairs as a function of weak label bins for two versions of the model. For example, for the set of pairs with weak labels within [0.0,0.2], the average relevance score obtained from embeddings with and without weak supervision are 0.12 and 0.24, respectively. As it can be seen, the relevance scores obtained from PHAN with supervision are more correlated with the weak labels than the ones obtained from PHAN without supervision. In fact, by employing the idea of supervised attention, we guide the model to find embeddings in a way that the relevance of hashtags and representative nodes become more aligned with the weak labels. Note that we do not claim that weak labels are fully accurate in determining the relevance scores; however, we suggest they can be beneficial in better recognizing the pattern, and we use the parameter λ to adjust their involvement level in the learning process.

4.4 Scale-freeness of the network (RQ3)

In this experiment, we study the validity of our basic assumption: What proportion of links are received by representative nodes? Figure 6 depicts the degree distribution of the user's links towards the entire set of nodes (raw) and towards representative nodes (filtered) in our Twitter and Weibo datasets. The set of representative nodes were selected by setting L to 120 and 10 in Twitter and Weibo datasets, respectively. As it can be seen, the out-degree towards representative nodes tends to be smaller than raw out-degree; however, a large percentage of users have out-degree larger than 10 towards representative nodes in both of the datasets. In other words, the idea of user profiling based on representative nodes is effective for a large proportion of users.

It should be noted that the proposed model can be further developed to become effective even for those users who do not have a

RIGHTSLINK

sufficient number of links towards representative nodes. For such users, as an alternative solution, we can rely on ordinary (non-representative) neighbors. That is, the target user's ordinary neighbors can be embedded based on their links towards representative nodes. Again, based on scale-free property, ordinary-neighbors are expected to have abundant links towards hub-nodes, which enables us to profile them based representative nodes. And next, the target user can be profiled based on the profiles of the ordinary neighbors. Indeed, we still rely on hub-nodes; however, we use ordinary-users as bridges to reach representative nodes.

5 RELATED WORKS

We review the studies that are related to our work. We first describe recent literature on the problem of Hashtag Recommendation followed by literature related to Attention Mechanism.

Hashtag Recommendation: Although hashtag-recommendation for *users* is an important task for user engagement in Microblogging website, it has not received much attention in the literature. To the best of our knowledge, [1] is the only work that attempts to address the problem. The model feeds the user-hashtag interaction matrix to a classic Matrix Factorization model to make recommendations. Clearly, the effectiveness of the model is quite limited, especially for users with zero or sparse interactions.

While our focus is on graph data (UHR task), the proposed model (when graph and text data are combined) can serve as a personalized hashtag recommendation model for Microblogs. The existing model for MHR can be categorized into two groups: 1) Traditional global models, 2) Personalized models

1) Vast majority of existing models for *hashtag recommendation for microblogs* address the task by only analyzing the content of the target Microblog. For example, classic methods such as tf-idf based retrieval models and translation based methods have been adopted to address the task [40][14] [15]. Deep Learning based hashtag recommendation models have also recently emerged. [37], [18], [6] and [13].

2) Some works also have tried to improve classic text based models by making personalized recommendations. Zhang *et al.* extended upon the translation based methods to leverage personal factors [41]. Analogously, in [16], user related features including tweeting-histories, location, and social influence were incorporated as auxiliary information to develop a personalized model. Also, recently personalized hashtag recommendation models for multimedia microblogs has been introduced [36]. Rawat et al. considered the contextual preferences of users on images (e.g. time and geo-location) to build user representation and use it in hashtag recommendation for images [32]. Park et al. build a model of user preferences based on their most frequently used hashtags and devised a Context Sequence Memory Network for hashtag recommendation for images based on user preferences and image features. [10].

While these works mainly focus on users with dense interaction history, our model can cover all of the users including those with sparse interactions by modeling them solely based on graph data.

Attention Mechanism: In general, to compute an upper-layer of the neural network, the attention mechanism can be used to determine a weight to each position in a lower-layer [3]. Attention-based models have been successfully applied for a wide range of sequence based tasks including text entailment [33], image captioning [39], sentence summarization [34] and recommender system [9, 20]. For example, a recent successful work in graph mining context has introduced an attention-based architecture to perform node classification of graph-structured data where the hidden representations of each node in the graph is computed by attending over its neighbors [35]. Also, a recent work in recommender systems has introduced an attention network which is capable of distinguishing which historical items in a user profile are more important for a prediction [20].

Classically, for models with low complexity and in presence of abundant training data, the attention components can be trained in an implicit manner based on the final output of the model [20, 30, 35]. However, that is not the case for model with higher complexity. In fact, some recent works in computer vision and language translation have shown that the accuracy of the attention map generated by an attention model learned in an implicit manner is not ensured [10, 11, 31]. For example, [11] shows there is a substantial gap in quality between the alignments obtained from the classic attention models and the human labeled alignments. To tackle this issue, in computer vision, explicit supervision has been injected to attention models using labeled data specifically acquired for the attention model [9]. We follow the same core idea in our attention model, however, we devise the notion of weakly supervision based on statistical models.

6 CONCLUSIONS

In this paper, we studied the problem of User-Hashtag Recommendation to facilitate the efficient use of hashtags on microblogging websites. Applying classic recommendation models to address the problem based on content data faces major drawbacks due to sparsity and ambiguity of content data. In fact, most of users are not active in terms of generating content/text data. Our key idea was to profile users' interest based on graph data and leverage it in the recommendation process. To efficiently exploit graph data, we introduce the notion of representative nodes and suggest that the embeddings of users in scale-free networks can be derived from these nodes which to our knowledge is the first instance to do so in graph mining context. We further propose a weakly supervised attention mechanism to perform a target/hashtag aware aggregation of representative nodes. Extensive experiments on two real-world datasets obtained from Twitter and Weibo websites verified the effectiveness of our model.

ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation IIS 16-19302 and IIS 16-33755, Zhejing University ZJU Research 083650, Futurewei Technologies HF2017060011 and 094013, UIUC OVCR CCIL Planning Grant 434S34, UIUC CSBS Small Grant 434C8U, and Advanced Digital Sciences Center Faculty Grant. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the author(s) and do not necessarily reflect the views of the funding agencies.

RIGHTSLINK

REFERENCES

[1] Hamidreza Alvari. 2017. Twitter hashtag recommendation using matrix factorization. *arXiv preprint arXiv:1705.10453* (2017).

[2] Sofía Aparicio, Javier Villazón-Terrazas, and Gonzalo Álvarez. 2015. A model for scale-free networks: application to twitter. *Entropy* 17, 8 (2015), 5848–5867.

[3] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. *international conference on learning representations* (2015).

[4] Albert-László Barabási and Eric Bonabeau. 2003. Scale-free networks. *Scientific american* 288, 5 (2003), 60–69.

[5] Nicola Barbieri, Francesco Bonchi, and Giuseppe Manco. 2014. Who to follow and why: link prediction with explanations. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 1266–1275.

[6] Nada Ben-Lhachemi and El Habib Nfaoui. 2018. Using tweets embeddings for hashtag recommendation in Twitter. *Procedia Computer Science* 127 (2018), 7–15.

[7] Jesús Bobadilla, Fernando Ortega, Antonio Hernando, and Abraham Gutiérrez. 2013. Recommender systems survey. *Knowledge-based systems* 46 (2013), 109–132.

[8] Hong-Ming Chen, Ming-Hsiu Chang, Ping-Chieh Chang, Min-Chun Tien, Winston H. Hsu, and Ja-Ling Wu. 2008. SheepDog: group and tag recommendation for flickr photos by automatic search-based learning. In *ACM Multimedia*.

[9] Jingyuan Chen, Hanwang Zhang, Xiangnan He, Liqiang Nie, Wei Liu, and Tat-Seng Chua. 2017. Attentive collaborative filtering: Multimedia recommendation with item-and component-level attention. In *Proceedings of the 40th International ACM SIGIR conference on Research and Development in Information Retrieval*. ACM, 335–344.

[10] Cesc Chunseong Park, Byeongchang Kim, and Gunhee Kim. 2017. Attend to you: Personalized image captioning with context sequence memory networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 895–903.

[11] Abhishek Das, Harsh Agrawal, Larry Zitnick, Devi Parikh, and Dhruv Batra. 2017. Human attention in visual question answering: Do humans and deep networks look at the same regions? *Computer Vision and Image Understanding* 163 (2017), 90–100.

[12] Dmitry Davidov, Oren Tsur, and Ari Rappoport. 2010. Enhanced Sentiment Learning Using Twitter Hashtags and Smileys. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters (COLING '10)*. Association for Computational Linguistics, 241–249.

[13] Kuntal Dey, Ritvik Shrivastava, Saroj Kaushik, and L. Venkata Subramaniam. 2017. EmTaggeR: A Word Embedding Based Novel Method for Hashtag Recommendation on Twitter. In *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*. 1025–1032.

[14] Zhuoye Ding, Xipeng Qiu, Qi Zhang, and Xuanjing Huang. 2013. Learning Topical Translation Model for Microblog Hashtag Suggestion.. In *IJCAI*. 2078–2084.

[15] Zhuoye Ding, Qi Zhang, and Xuanjing Huang. 2012. Automatic hashtag recommendation for microblogs using topic-specific translation model. *Proceedings of COLING 2012: Posters* (2012), 265–274.

[16] Wei Feng and Jianyong Wang. 2014. We can learn your# hashtags: Connecting tweets to explicit topics. In *2014 IEEE 30th International Conference on Data Engineering*. IEEE, 856–867.

[17] Yuyun Gong and Qi Zhang. 2016. Hashtag Recommendation Using Attention-based Convolutional Neural Network. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI'16)*. AAAI Press, 2782–2788.

[18] Yuyun Gong and Qi Zhang. 2016. Hashtag Recommendation Using Attention-Based Convolutional Neural Network.. In *IJCAI*. 2782–2788.

[19] Xiangnan He, Tao Chen, Min-Yen Kan, and Xiao Chen. 2015. TriRank: Review-aware Explainable Recommendation by Modeling Aspects. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management (CIKM '15)*. ACM, 1661–1670.

[20] Xiangnan He, Zhenkui He, Jingkuan Song, Zhenguang Liu, Yu-Gang Jiang, and Tat-Seng Chua. 2018. NAIS: Neural Attentive Item Similarity Model for Recommendation. *IEEE Transactions on Knowledge and Data Engineering* (2018).

[21] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. 2017. Neural Collaborative Filtering. In *Proceedings of the 26th International Conference on World Wide Web (WWW '17)*. International World Wide Web Conferences Steering Committee, 173–182.

[22] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation* 9, 8 (1997), 1735–1780.

[23] Liangjie Hong, Amr Ahmed, Siva Gurumurthy, Alexander J. Smola, and Kostas Tsioutsiouliklis. 2012. Discovering Geographical Topics in the Twitter Stream. In *Proceedings of the 21st International Conference on World Wide Web (WWW '12)*. ACM, 769–778.

[24] Haoran Huang, Qi Zhang, Yeyun Gong, and Xuanjing Huang. 2016. Hashtag Recommendation Using End-To-End Memory Networks with Hierarchical Attention. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*.

[25] Haoran Huang, Qi Zhang, Xuanjing Huang, et al. 2017. Mention recommendation for Twitter with end-to-end memory network. In *Proc. IJCAI*, Vol. 17. 1872–1878.

[26] Amin Javari, HongXiang Qiu, Elham Barzegaran, Mahdi Jalili, and Kevin Chen-Chuan Chang. 2017. Statistical Link Label Modeling for Sign Prediction: Smoothing Sparsity by Joining Local and Global Information. In *2017 IEEE International Conference on Data Mining (ICDM)*. IEEE, 1039–1044.

[27] Ralf Krestel, Peter Fankhauser, and Wolfgang Nejdl. 2009. Latent Dirichlet Allocation for Tag Recommendation. In *Proceedings of the Third ACM Conference on Recommender Systems (RecSys '09)*. ACM, 61–68.

[28] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. 2010. What is Twitter, a social network or a news media?. In *Proceedings of the 19th international conference on World wide web*. AcM, 591–600.

[29] Lemao Liu, Masao Utiyama, Andrew M. Finch, and Eiichiro Sumita. 2016. Neural Machine Translation with Supervised Attention. *international conference on computational linguistics* (2016), 3093–3102.

[30] Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. 2017. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Vol. 6. 2.

[31] Tingting Qiao, Jianfeng Dong, and Duanqing Xu. 2018. Exploring human-like attention supervision in visual question answering. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

[32] Yogesh Singh Rawat and Mohan S Kankanhalli. 2016. ConTagNet: Exploiting user context for image tag recommendation. In *Proceedings of the 24th ACM international conference on Multimedia*. ACM, 1102–1106.

[33] Tim Rocktäschel, Edward Grefenstette, Karl Moritz Hermann, Tomáš Ko iský, and Phil Blunsom. 2016. Reasoning about Entailment with Neural Attention. *international conference on learning representations* (2016).

[34] Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. A Neural Attention Model for Abstractive Sentence Summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. 379–389.

[35] Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Li'o, and Yoshua Bengio. 2018. Graph Attention Networks. In *Proceedings of ICLR 2018, the 6th International Conference on Learning Representations*.

[36] Yinwei Wei, Zhiyong Cheng, Xuzheng Yu, Zhou Zhao, Lei Zhu, and Liqiang Nie. 2019. Personalized Hashtag Recommendation for Micro-videos. *arXiv preprint arXiv:1908.09987* (2019).

[37] Jason Weston, Sumit Chopra, and Keith Adams. 2014. # tagspace: Semantic embeddings from hashtags. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 1822–1827.

[38] Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and Philip S Yu. 2019. A comprehensive survey on graph neural networks. *arXiv preprint arXiv:1901.00596* (2019).

[39] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. 2016. Image captioning with semantic attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4651–4659.

[40] Eva Zangerle, Wolfgang Gassler, and Gunther Specht. 2011. Recommending#-tags in twitter. In *Proceedings of the Workshop on Semantic Adaptive Social Web (SASWeb 2011). CEUR Workshop Proceedings*, Vol. 730. 67–78.

[41] Qi Zhang, Yeyun Gong, Xuyang Sun, and Xuanjing Huang. 2014. Time-aware personalized hashtag recommendation on social media. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*. 203–212.

[42] Yizhou Zhang, Yun Xiong, Xiangnan Kong, Shanshan Li, Jinhong Mi, and Yangyong Zhu. 2018. Deep Collective Classification in Heterogeneous Information Networks. In *Proceedings of the 2018 World Wide Web Conference (WWW '18)*. International World Wide Web Conferences Steering Committee, 399–408.

RIGHTSLINK