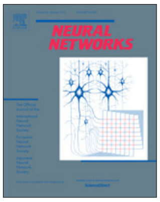
Neural Networks 106 (2018) 152–167

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet



Spatial context-aware user mention behavior modeling for mentionee recommendation

Kai Wang[a], Weiyi Meng[b], Jiang Bian[c], Shijun Li[a,*], Sha Yang[d]

[a] *Computer School, Wuhan University, Wuhan 430072, China*
[b] *Department of Computer Science, Binghamton University, Binghamton 13902, USA*
[c] *Department of Health Outcomes and Biomedical Informatics, University of Florida, Gainesville 32611, USA*
[d] *School of Computer Science and Technology, Hankou University, Wuhan 430212, China*

ARTICLE INFO

Article history:
Received 12 March 2018
Received in revised form 25 May 2018
Accepted 11 July 2018
Available online 18 July 2018

Keywords:
Mention behavior modeling
Spatial context awareness
Joint probabilistic model
Mentionee recommendation

ABSTRACT

As one of the most common user interactive behaviors in many social media services, *mention* plays a significant role in both user interaction and information cascading. While an increasing line of work has focused on analyzing the mention mechanism for information diffusion, the essential problem of *mentionee recommendation* from the perspective of common users, i.e., how to find mentionees (mentioned users) who are most likely to be notified by a mentioner (mentioning user) for *knowing* a post, has been seldom investigated. This paper aims to develop personalized recommendation techniques to automatically generate mentionees when a user intends to mention others in a post. After analyzing real-world social media datasets we observe that users' mention behaviors are influenced by not only the semantic but also the spatial context factors of their mentioning activities, which motivate the needs for spatial context-aware user mention behavior modeling. In light of these, we proposed a joint probabilistic model, named Spatial COntext-aware Mention behavior Model (SCOMM), to simulate the process of generating users' location-tagged mentioning activities. By exploiting the semantic and spatial context factors in a unified way, SCOMM was able to reveal users' preferences behind their mention behaviors and provide a knowledge model for accurate mentionee recommendations. Furthermore, we designed an Item-Attribute Pruning (IAP) algorithm to overcome the curse of dimensionality and facilitate online top-*k* query performance. Extensive experiments were conducted on two real-world datasets to evaluate the performance of our methods. The experimental results demonstrated the superiority of our approach by making more effective and efficient recommendations compared with other state-of-the-art methods.

1. Introduction

With the rising popularity of social media, it has been easier for people to share information in various forms online. In addition to message sharing, users may also adopt different types of behaviors to interact with each other. For example, users of Twitter[1] and Weibo[2] can *reply* and *retweet* posts, mark *hashtags mention* others for noticing, as well as *follow* and *subscribe* to interesting users. Among these online user interactive behaviors, *mention* is playing an increasingly significant role in both user interaction and information cascading due to its unique advantages in alleviating information overload issues (Bobadilla, Ortega, Hernando, & Gutiérrez, 2013). Nowadays, social media hold a huge and growing number of active users. In late 2017, there were 330 million monthly active users on Twitter, who were publishing 500 million tweets per day, while the average number of friends per user was 707.[3] Hence, when people intend to mention others in a post, a small list of suggested mentionees would definitely be helpful. Currently, the solution given by social media platforms like Twitter, Weibo and Facebook[4] is to provide mentionee suggestions after a user inputs the mentioning symbol "@". However, the suggestions given by these services are usually based on either completion from partial inputs or users' mentioning histories, which usually lead to less desirable.

Although several approaches have been proposed to tackle this user recommendation problem (Li, Song, Liao, & Liu, 2015; Shen, Ding, Qiao, Cheng, & Wang, 2016; Wang et al., 2013; Zhou et al., 2015), most of them have focused only on the information

* Correspondence to: Room B507, Computer School, Wuhan University, Wuhan 430072, China.
E-mail address: shjli@whu.edu.cn (S. Li).

[1] http://www.twitter.com/.

[2] http://www.weibo.com/.

[3] http://about.twitter.com/company/.

[4] http://www.facebook.com/.

https://doi.org/10.1016/j.neunet.2018.07.007
0893-6080/© 2018 Elsevier Ltd. All rights reserved.

propagation aspect of users' mention behaviors, e.g., to find users who can expand the diffusion of a post. In fact, mention functions both as a communication tool and an information propagation channel (Jiang, Sha, & Wang, 2015; Tang, Ni, Xiong, & Zhu, 2015). Intuitively, the communication aspect of the mention is more valuable to common users than to the relatively few media workers and advertisers in social media. For a common user, the spreading speed or the popularity of a post may not be the primary consideration when the mention activity happens. In other words, the purpose of mentioning for a common users is more likely to notify mentionees to *know* about a post rather than to spread it. Besides, recent studies (Kogan, Palen, & Anderson, 2015; Lee, Mahmud, Chen, Zhou, & Nichols, 2014) suggested that information diffusion relies more on retweets than mentions in social media. Hence in this work, we investigate the problem of mentionee recommendation by analyzing and modeling user mention behaviors from the perspectives of common users. Different from the previous works, the aim of user mention behavior we focus on is not limited to the spreading of a post but its generalized attributes of users' online interactions.

While it was well investigated that the textual content information is critical to a user's mentionee selection in a mentioning instance (Chen et al., 2012; Gong, Zhang, Sun, & Huang, 2015; Weng, Lim, Jiang, & He, 2010; Zhou et al., 2015), the effect of other contextual factors on mentionee selection has seldom been investigated. Recently, triggered by the widespread adoption of GPS-enabled tagging of user generated records via mobile devices and online social media services, the diffusion of the location dimension in online social media provides us valuable opportunity to better understand the correlation between users' online behaviors and their physical movements. In this case, we can capture users' behavioral preferences more accurately by incorporating their spatial factors in the physical world. In this work, after analyzing two large real-world social media datasets, we observe that users' mention behaviors are influenced by not only the semantic but also the spatial context factors of their mentioning activities. More specially, we observe that the home locations of mentionees for a mentioner tend to be within a specific geographic area, and the mentioners are more likely to select mentionees who live close to their standing locations (see Section 3.2). These observations reveal the geographical relationships between mentioners and mentionees, which further motivate the need for spatial context-aware user mention behavior modeling and can be exploited for mentionee recommendations.

The problem studied in this article is very challenging, however. First, the recommendation scenario of this task should be highly personalized since users have different preferences. It is a challenging work to infer user preferences precisely from the massive and knowledge-sparse social media data. The problem is more severe when considering the sparsity of user-word matrix in social media services (Yin, Cui, Zhou, Wang, Huang, & Sadiq, 2016). Second, social media users are allowed to select mentionees beyond their neighborhood, which means anyone can be the candidate and leads to an extremely large candidate space. Third, mentionee suggestion services work in an online manner and users are used to receive suggestions right after they enter the symbol "@". Hence, efficiency is also an essential requirement for this task. In other words, we need to quickly find the right results in an extremely large candidate space.

In this work, we aim to tackle the mentionee recommendation problem, i.e., recommend mentionees when a user intend to mention others in a post, from the perspective of common users' online interactions. Specifically, we investigate the problem by exploring users' online mention behaviors. A joint latent-class probabilistic model, named Spatial COntext aware Mention behavior Model (SCOMM), is proposed to simulate the process of generating users' mentioning activities by synthetically considering both semantic and spatial context factors in a unified way. As users' mention behaviors are influenced by both semantic and spatial context factors, two key latent features are proposed in SCOMM –semantic topic and geographical area – which are responsible for generating semantic and geographical attributes of users' mentioning activities, respectively. Based on these two features, SCOMM simultaneously learns and models the semantic patterns of mentioners, the geographical clustering areas of mentionees, and their joint effects on mentioners' movement patterns. After training the SCOMM offline to obtain a knowledge model containing the necessary insights about users' mention behaviors, we retrieve the top-k mentionees for an online mentioning post. To support efficient online query processing, we designed an Item-Attribute Pruning (IAP) algorithm to prune the searching space in both spatial and semantic dimensions simultaneously.

Our major contributions can be summarized as follows:

- We investigated the problem of mentionee recommendation from the perspective of common users and statistically studied the geographical factors that affect users' mentionee selections. As far as we know, this work is the first to study the impact of the users' spatial context factors on their mention behaviors.
- We proposed a probabilistic approach to model the semantic-aware and spatial-aware user mention behaviors comprehensively, which strategically incorporates both content and spatial analysis techniques.
- We designed an efficient online pruning algorithm to speed up top-k mentionee recommendations by pruning the searching space in both spatial and semantic dimensions simultaneously.
- Through extensive experiments constructed on large real-world datasets, we demonstrated the effectiveness and efficiency of our proposed approaches.

The rest of this paper is organized as follows. Section 2 surveys the existing work related to our research. Section 3 describes our proposed model in detail. In Section 4, we deploy our proposed model for efficient mentionee recommendation. We conduct a series of experiments using data collected from real-world social media sites for evaluation in Section 5. Conclusion is presented in Section 6.

2. Related work

Our work is related to three research threads: (1) user interactive behavior analysis and modeling on social web; (2) spatial context-aware topic modeling and (3) user recommendation on social media.

2.1. Social user interactive behavior analysis

User interactive behavior analysis plays an important role in social media user profiling and personality mining. It sheds light on the correlations between what people think and what people act. With the dramatic increase of users, social media services are offering more and more online functions to support users' interactions, which motivated an increasing line of work focusing on user interactive behavior analysis. For example, Xu, Zhang, Wu, and Yang (2012) modeled user posting behavior based on the observation that users' posting behaviors are mainly influenced by breaking news, posts from friends and their intrinsic interests. Qiu, Zhu, and Jiang (2013) investigated users' interactions with their generated text, and proposed an latent probabilistic method to model users' topic interests and behavioral patterns. Yin, Cui, Chen, Hu, and Zhou (2015) observed that users' rating behaviors are significantly influenced by the intrinsic interests of the user and the attention of the general public. In recent years, retweeting behavior has attracted much interest since it provides a strong indication of the direction of information flow in social media services (Bi & Cho, 2016; Chen et al., 2012; Kogan et al., 2015; Lee et al., 2014). For example, Chen et al. (2012) constructed a personalized tweet recommendation model through analyzing users' retweeting behavior; Bi and Cho (2016) proposed two Bayesian nonparametric models to identify users' personal interests by incorporating users' retweet content and their retweet behavior. In addition to those, attentions have also been paid on mentioning (Tang et al., 2015; Wang et al., 2013; Zhang, Wang, Yin, Wang, & Yu, 2017; Zhou et al., 2015) and hashtagging (Alam, Ryu, & Lee, 2017; Ding, Qiu, Zhang, & Huang, 2013). In our work, with a different purpose, we conducted a user mention behavior analysis in social media services to reveal user preferences and provide a knowledge model for mentionee recommendation.

2.2. *Spatial context aware topic modeling*

Topic models provide us an unsupervised and useful tool to analyze not only content but many types of discrete data. There are many variants of topic models such as Probabilistic Latent Semantic Analysis (PLSA) (Hofmann, 1999) and Latent Dirichlet Allocation (LDA) (Blei, Ng, & Jordan, 2003). In fact, our proposed SCOMM can also be classified into the family of LDA-based models. These variants have been proven to be helpful in discovering spatial-topical structures from large user-generated datasets. Recently, some of the works focus on mining geographical knowledge by studying the interaction between users, locations and topics in social network (Fang, Xu, Hossain, & Muhammad, 2016; Hu & Ester, 2013; Kurashima, Iwata, Irie, & Fujimura, 2010; Liu, Fu, Yao, & Xiong, 2013; Wang, Yin et al., 2015; Yin & Cui, 2016; Yin, Sun, Cui, Hu, & Chen, 2013). For example, Kurashima et al. (2010) proposed a probabilistic behavior model which combined topic models and Markov models to estimate the probability of a photographer visiting a location. Hu and Ester (2013) explored topic modeling considering the spatial and textual aspects of user posts to predict future user locations. Yin et al. (2013) proposed a location-content-aware generative model that quantifies and incorporates both local preference and item content in the spatial item recommendation task. Similarly, the geographical probabilistic factor analysis framework designed by Liu et al. in Liu et al. (2013) strategically took into account various factors to capture the geographical influences on a user's check-in behavior. More recently, Fang et al. proposed a topic model based spatial–temporal context-aware location recommendation system in Fang et al. (2016) to offer an individual user a set of location items within a geographical range. Different from them, our work focused on inferring users' interactive behavioral preferences by considering both the semantic and the spatial context information of users' behavioral activities.

2.3. *User recommendation on social media*

User recommendations have become a rich research area within the broad recommender community and social recommendations in particular. Extensive work has been undertaken to a variety of user recommendation problems on social media platforms such as friend and followee recommendation (Gupta et al., 2013; Hannon, Bennett, & Smyth, 2010), expert recommendation (Cheng, Caverlee, Barthwal, & Bachani, 2014; Ge, Caverlee, & Lu, 2016) and mention related recommendation (Tang et al., 2015; Wang et al., 2013; Zhang et al., 2017; Zhou et al., 2015). Specially, the mentionee recommendation problem has been studied from different aspects. For example, Wang et al. (2013) trained a machine learned ranking function with features such as user interest match and user influence to solve the "whom to mention" problem. The main goal of their work was to make a tweet spread more quickly and widely. Zhou et al. (2015) studied the mentionee ranking problem from the perspective of information overload. They proposed a mentionee ranking model to tackle the problem by exploring features such as content influence and user affinity. The CAR recommendation framework developed by Tang et al. in Tang et al. (2015) is a context-aware approach aiming to find the right target audiences who will respond to a promotion-oriented post. They employed the Ranking Support Vector Machine (SVM) model to resolve the ranking based mentionee suggestion problem after related features are extracted. The goal of their work is to achieve high response rate. More recently, Gong et al. proposed A-UUTTM (Gong et al., 2015), a topical translation model incorporating the current content and the histories of mentionees to perform the task, which was the state-of-the-art approach used for mentionee recommendation.

The distinction between existing methods and our approach can be summarized as follows. First, instead of focusing on how to make a tweet spread more quickly or get a higher response rate, our work is motivated by the essential problem of mentionee recommendation for common users, i.e., find mentionees who are most likely to be noticed for *knowing* a post. The aim of user mention behavior we focused on is not limited to the spread of a post but its generalized attributes of users' online interactions. Second, instead of solely leveraging the textual content, we incorporated the geographical data of both mentioners and mentionees. Experimental results have demonstrated that the spatial context information of users' mentioning activities is very helpful for the performance improvement of mentionee recommendation system. Third, instead of producing offline recommendations in a very time-consuming manner, we aimed to provide a practical recommendation system to quickly answer online queries. To achieve that, we designed an efficient pruning algorithm to facilitate online top-k query performance.

3. Spatial context aware joint mention behavior model

In this section, we first introduce the notations and concepts used in this paper and then formulate the problem definition. To illustrate the proposed SCOMM, we first present several observations and intuitions behind the modeling process and then detail our model, followed by the model inference.

3.1. *Preliminary*

Table 1 shows the notations used in this paper. The key concepts are defined as follows.

Definition 1 (*Mentioner and Mentionee*). We denote a user u as a mentioner if she has mentioned others (i.e., mentionees) in at least one mentioning post. Clearly, "Mentioner" and "mentionee" are relative concepts in this paper, i.e., user can be both a mentioner in one post and a mentionee in another. In our model, a mentionee has two attributes: an identifier m and a home location l_m. We use U and M to denote the sets of all mentioners and mentionees, respectively.

Following the recent works of Li, Wang, Deng, Wang, and Chang (2012), Yin et al. (2016), Yin and Cui (2016), we define a mentionee's home location as the place where she lives in geographical coordinates. Usually, Social media users can choose whether to make their home location publicly available. For users whose home locations were not explicitly given, we adopted a simple and effective ground truth user home location acquiring method (see Section 5.1.1).

Table 1
Notations in our work.

Variable	Description
z, w, m, a, l_m, l_d	Index of topic, word, mentionee, area, home location of m, and current location of mentioning activity d, respectively
T, A, U, M, W	The sets of topics, areas, mentioners, mentionees and words, respectively
K, Q, N, E, V	The numbers of topics, areas, mentioners, mentionees and words, respectively
ϕ_u	Multinomial distribution over areas, representing the distribution of the mentionee clustering areas of mentioner u
θ_u	Multinomial distribution over topics, representing the semantic pattern of u
ϑ_z, ϑ_b	Multinomial distribution over words for topic z, and background multinomial distribution over words, respectively
$\psi_{z,w}$	Multinomial distribution over mentionees specific to topic z and word w
$\pi_{z,a}$	Multinomial distribution over user standing locations specific to topic z and area a
μ_a, Σ_a	Mean vector and covariance matrix of area a
y	A switch that determines which distribution the word is generated from, y = 0 or 1
$\alpha, \gamma, \eta, \xi$	Dirichlet priors to multinomial distributions θ_u, $\psi_{z,w}$, ϕ_u and $\pi_{z,a}$, respectively
β	Dirichlet priors to multinomial distributions ϑ_z and ϑ_b
ρ, λ	Bernoulli distribution to generate switch y, and Beta prior to ρ (denoted by $\lambda = \{\lambda_{1,} \lambda_2\}$), respectively

Definition 2 (*Location-Tagged Mentioning Activity*). In this work, a location-tagged mentioning activity d is represented by a 4-tuple (u, W_d, l_d, M_d) denoting that a mentioner u mentions mentionees M_d with words W_d at the location l_d. Note that M_d denotes a set of mentionees since a user may mention multiple mentionees in a single post. We use L_{M_d} to denote the set of the home locations of the mentionees in M_d.

Definition 3 (*Mentioning Document*). For each mentioner u, the mentioning document D_u is the set of location-tagged mentioning activities of u. The dataset D consists of all the mentioning documents. i.e., $D = \{D_u | u \in U\}$. L_D denotes the set of attached locations with mentioning activities, i.e., $L_D = \{l_d | d \in D\}$, and L_M represents the set of home locations of all mentionees, i.e., $L_M = \left\{L_{M_d} | d \in D\right\}$.

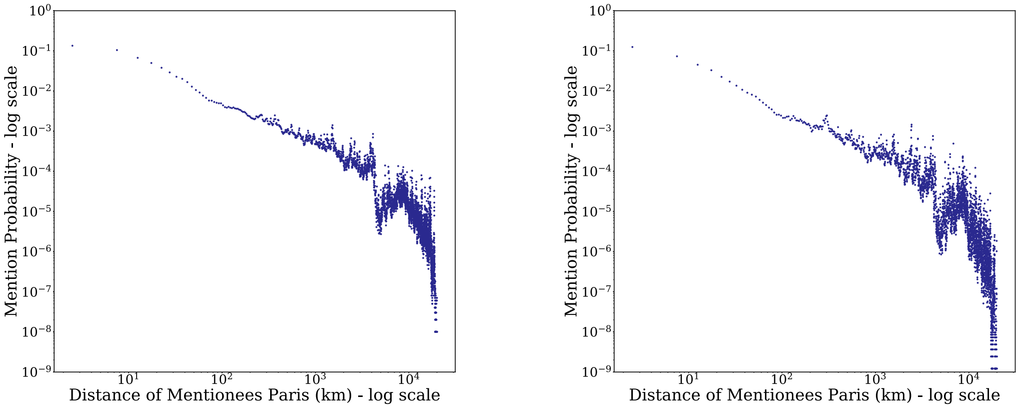
Given a dataset D as a collection of user mentioning documents, we aim to provide mentionee recommendations for social media users. We formally define the problem as follows.

Definition 4 (*Mentionee Recommendation Problem*). For a querying post published by u_q with words W_q at her current location l_q, i.e., the query is $q = \left\{u_q, l_q, W_q\right\}$, our goal is to recommend a set of top-k mentionees M_q that u_q is most likely to mention.

3.2. Observations and intuitions behind SCOMM

In this section, we introduce the observations and intuitions behind the modeling process of SCOMM. Specifically, we performed a statistical analysis on two large real-world datasets collected from Weibo and Twitter. Each of these datasets contains over 550,000 location-tagged mentioning posts published by more than 130,000 mentioners and over 180,000 related mentionees. Two interesting findings are observed as follows.

1. **Geographical clustering of mentionees**. We observed a geographical clustering phenomenon which suggests that the home locations of the mentionees mentioned by the same mentioner tend to be geographically clustered. Inspired by Ye et al. in Ye, Yin, Lee, and Lee (2011), we calculated the geographical distances between all pairs of home locations of mentionees who were mentioned by the same mentioner, and then obtained the probabilities of mentionee pairs being mentioned over their geographical distances. As shown in Fig. 1, the mention probabilities appear to follow a power-law distribution, and a significant percentage of mentionee pairs appears to be within short geographical distances. This phenomenon can be attributed to the geographical influences on users' mention behaviors which may be intuitively explained by the following tendencies. First, work from the computational social science community has established that online social ties are often formed over short physical distances (Takhteyev, Gruzd, & Wellman, 2012). Empirical studies such as McGee, Caverlee and Cheng (2011, 2013) also suggested that most of the users' online friends are geographically local. These investigations revealed that social users primarily communicate over short geographic distances, which leads to more frequent online interactions between users in short distances. This result is in agreement with previous social media studies (Compton, Jurgens, & Allen, 2014; Jurgens, 2013). Second, the evolution of location-centric online communities in social services accelerates the online clustering of users who are physically close (Brown, Nicosia, Scellato, Noulas, & Mascolo, 2012; Lim, Chan, Leckie, & Karunasekera, 2015; Takhteyev et al., 2012). People in the same location-centric online communities shared similar interests within certain geographical areas. Nevertheless, in certain cases, one may be interested in



(a) Twitter

(b) Weibo

Fig. 1. Log–log histogram of the geographical probability distribution of mentionees pairs.

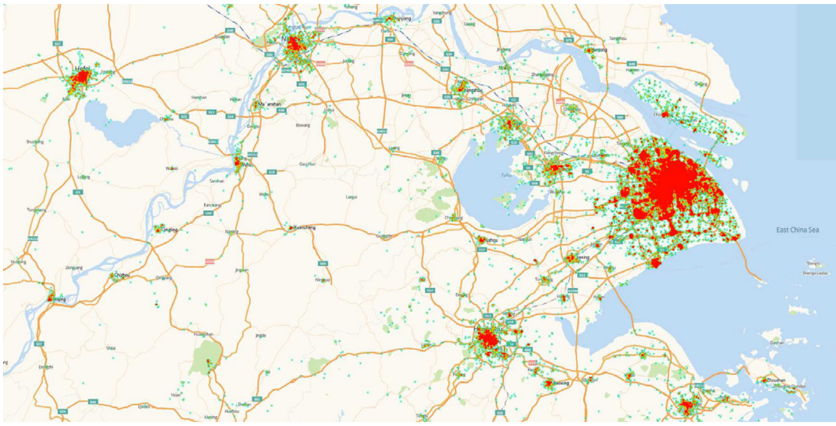


Fig. 2. The distribution of mentioners' standing locations for mentionees who live in Shanghai.

new location-centric online communities, even the centric locations are far from home.

2. **Geographical approximation of mentioning**. Another interesting observation is that mentioners tend to select mentionees who were close to their standing locations. For example, we drew mentionees whose home location are located in Shanghai, and then plotted a heat map to depict the geographical distribution of posts with respect to these mentionees (i.e., the standing location distribution of mentioners), as shown in Fig. 2. According to our survey, 34.53% of posts were exactly located in Shanghai, and 61.21% of posts were located in Shanghai and its surrounding provinces (i.e., Jiangsu and Zhejiang) as Fig. 2 depicts. This is much higher than the probabilities in other areas (e.g., 3.42% in Beijing and 7.35% in Guangdong province). This phenomenon may be subject to both "travel locality" (Scellato, Noulas, Lambiotte, & Mascolo, 2011) and "seeking novelty" (Lian et al., 2015; McGee et al., 2013; Wang, Yuan et al., 2015) of users. On the one hand, it echoes the above discussion about the correlation between physical proximity and user interaction tendency. On the other hand, one may be interested in exploring local places and interacting with local people when she travels to a new city.

In addition, there are two intuitions that motivates the modeling.

1. **Semantic patterns of mentioners**. Intuitively, interesting content is a very important factor to trigger online interactions. As investigated in many works (Chen et al., 2012; Lee et al., 2014; Weng et al., 2010; Zhou et al., 2015), the content-derived semantic pattern of a mentioner is critical to her decision-making process in the selection of mentionees. Empirical studies also suggested that to a specific mentionee, the related mentioning documents often exhibit a strong semantic regularity (Li et al., 2015; Tang et al., 2015). Thus, to our modeling, the contents of a mentioner's mentioning documents provide important clues to her selection of mentionees.
2. **Spatial patterns of mentioner semantics**. For Twitter-like social media services, users are allowed to add a Geo-tag (latitude and longitude) or a Point of Interest (POI) tag to a post to support the check-in functionality. Previous studies have demonstrated that the geographically proximate locations that users have checked in are more likely to share similar semantic characteristics and belong to the same semantic category (Ye, Shou, Lee, Yin, & Janowicz, 2011; Yin et al., 2016, 2013). In our task, the check-in information is embedded in users' location-tagged mentioning activities. As users' location-tagged mention behaviors can be seen as an extension of their check-in behaviors in terms of user interactions, the similar spatial patterns are also implied in their mentioning activities.

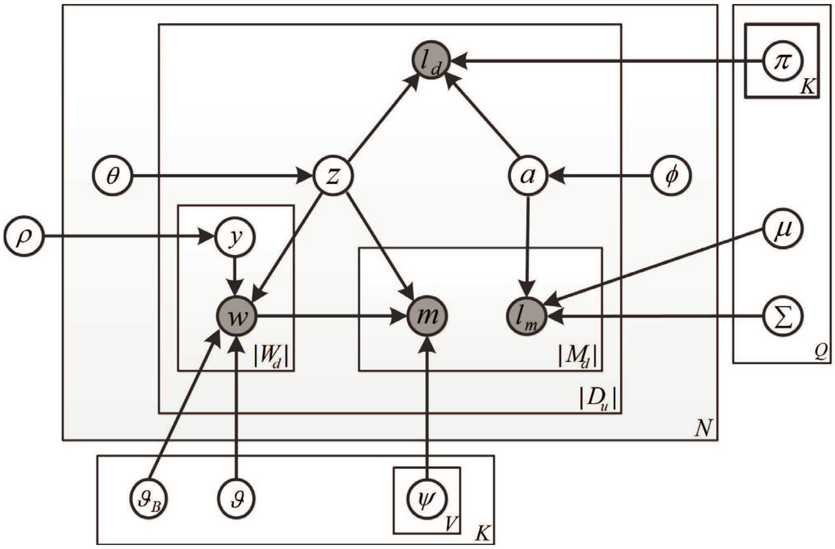


Fig. 3. The graphic representation of the proposed SCOMM. Priors are omitted.

3.3. Model structure

To infer the joint effects from content and spatial context information on users' mention behaviors, we propose a probabilistic model SCOMM to simulate the process of generating users' mentioning activities. Fig. 3 shows the graphical representation of SCOMM, where the observations, e.g., words, are shown as shaded circles, and the hidden variables e.g., topics are shown as unshaded ones. We use two key latent factors, topic z and area a, to generate the semantic attributes (e.g., words) and the spatial attributes (e.g., geographical coordinates) of users' mentioning activities, respectively. Based on these two factors, SCOMM jointly learns and models mentioner's semantic patterns, geographical clustering areas of mentionees, as well as their joint effects on mentioners' movement patterns.

Based on intuition 1, we adopted latent topics to characterize users' interests to overcome the sparsity of the user-word matrix, inspired by recent work on user interest modeling with textual content (Michelson & Macskassy, 2010; Zhao et al., 2011). Note that, the traditional topic models adopt a one-to-one topic-word assignment, i.e., they assign a topic label to each word (Blei et al., 2003). While this assumption makes sense when modeling long documents, it does not handle short documents well (Ding et al., 2013; Yin et al., 2016). Hence in SCOMM, for each mentioning activity d, we assumed that all the words in W_d share the same topic z, and each mentioner u holds a mixture of topics. Specifically, for each individual mentioner u, we inferred u's topical interest distribution over a set of topics (i.e., θ_u) according to the textual contents of u's mentioning activities. Thus, the quality of topics is critical to capturing users' semantic patterns. Motivated by the modeling mechanism of topical translation model (Ding et al., 2013; Gong et al., 2015; Huang et al., 2012), which inherits the advantages of both topic and translation models to efficiently capture the content information from short textual data, we followed the basic translation modeling framework and assumed that the generation of mentioning in the post can be regarded as a translation process from the textual words to mentionees' usernames. Hence, a topic z in SCOMM is not only responsible for generating words W_d but also for generating mentionees' usernames M_d. In other words, each topic z in SCOMM is not only associated with a distribution over words (i.e., ϑ_z) but also with a distribution over mentionees (i.e., $\psi_{z,w}$). Furthermore, A joint factor topic-word (z, w) was introduced to generate the usernames of mentionees according to the topical translation assumptions. Thus in SCOMM, each topic-word (z, w) is associated with a distribution over mentionee usernames (i.e., $\psi_{z,w}$). Through the above modeling processes, we can infer the semantic correlations among topics, mentionees and words.

Observation 1 indicates that mentionees tend to be geographically clustered, which reveals the physical correlations between users in online interactions. Hence, it is also necessary for SCOMM to capture this spatial factor by modeling the mentionees' clustering areas. In this work, we divided all the home locations of mentionees into R geographical areas according to their home location distributions. For each individual mentioner u, we apply a multinomial distribution over areas (i.e., ϕ_u) to model the distribution of the mentionee clustering areas of u. Since an area a should be coherent in the geographical space, we model a as a geographical Gaussian distribution using the continuous geographic locations (i.e., latitudes and longitudes), instead of a multinomial distribution over discrete locations that may be far from each other. Specifically, the location of mentionee m is characterized by $l_m \sim \mathcal{N}(\mu_a, \Sigma_a)$, where μ_a and Σ_a are the mean vector and covariance matrix of area a, respectively.

According to intuition 2, the geo-tagged locations (i.e., mentioners' standing locations) provide important clues about semantics of users mentioning activities. Besides, Observation 2 also indicates that there is a physical proximity correlation between a mentioner's standing location and the geographical distributions of her mentionees. Hence we can see that, a mentioner's movement pattern is influenced by the joint effects of her semantic interest and the geographical clustering areas of her mentionees. In this work, we use a joint latent factor topic-area to capture this joint effect. More specifically, the latent topic-area is responsible for generating the tagged locations of users' mentioning activities, and we associate each topic-area (z, r) with a distribution over mentioners' standing locations, i.e., $\pi_{z,a}$. This joint latent factor serves to seamlessly unify the modeling process of mentioner semantic pattern and mentionee clustering area, which also enables geographical clustering and topic modeling to influence and enhance each other under a unified framework.

3.4. The generative process of SCOMM

The generative process of a location-tagged mentioning activity d made by a mentioner u is based on the following assumptions. First, u selects a topic z according to her topical interest θ_u, then a

Algorithm 1: The generative Process of SCOMM

```
Draw $\rho \sim Beta(\lambda_1, \lambda_2)$, $\vartheta_b \sim Dirichlet(\beta)$;
for each $u \in U$ do
    Sample her distribution over topics $\theta_u \sim Dirichlet(.|\alpha)$;
    Sample her distribution over areas $\phi_u \sim Dirichlet(.|\eta)$;
for each topic $z \in T$ do
    Sample a distribution over words $\vartheta_z \sim Dirichlet(.|\beta)$;
    for each word $w \in W$ do
        Sample a distribution over topics and words
        $\psi_{z,w} \sim Dirichlet(.|\gamma)$;
    for each area $a \in A$ do
        Sample a distribution over topics and areas
        $\pi_{z,a} \sim Dirichlet(.|\eta)$;
for each $D_u \in D$ do
    for each mentioning activity $d \in D_u$ do
        Sample a topic indicator $z_d \sim Multi(.|\theta_u)$;
        for each token $w \in W_d$ do
            Toss a coin $y_w \sim Bernoulli(\rho)$;
            if $y_w = 0$ then
                Sample word $w \sim Multi(.|\vartheta_z)$;
            else
                Sample word $w \sim Multi(.|\vartheta_b)$;
        Sample an area indicator $a \sim Multi(.|\phi_u)$;
        Sample check-in location $l_d \sim Multi\left(.\left|\pi_{z,a}\right.\right)$;
        for each mentionee mentioned in d, $m \in M_d$ do
            Sample mentionee $m \sim P\left(.\left|z_d, W_d, \psi_{z,w}\right.\right)$;
            Sample $m$'s home location $l_m \sim \mathcal{N}(\mu_a, \Sigma_a)$;
```

sequence of words W is chosen according to either the topic-word distribution ϑ_z or the background word distribution ϑ_{back}. After the generation of the words, u selects several mentionees M according to topic z, words W and probability $P\left(.|z, W, \psi_{z,w}\right)$. Aside from the semantic content, u also needs to choose an area a according to the mentionee geographical distribution of u, i.e., ϕ_u. With the chosen a, the geographical coordinates of the home location of each mentionee $m \in M$ are generated according to the spatial distribution of a, i.e., $\mathcal{N}(\mu_a, \Sigma_a)$. At last, with the chosen topic z and area a, the tagged location l_d of d is generated according to the joint topic-area factor $\pi_{z,a}$. The generative story of SCOMM can be summarized as Algorithm 1.

3.5. Model inference

In this section, we describe how to learn the parameters of our model. In general, we need to obtain parameters that maximize the marginal log-likelihood of the observed random variables w, m, l_d and l_m. Accordingly, the marginalization is performed with respect to the latent random variables z and a. However, the marginal log-likelihood cannot be computed tractably due to the coupling among the latent variables. Therefore, we devise an approximate learning method based on collapsed Gibbs sampling (Griffiths & Steyvers, 2004) to maximize the complete data likelihood as shown in Eq. (1). Specifically, we assume that the priors follow symmetric Dirichlet, which are conjugate priors for multinational (or beta for Bernoulli). For simplicity, we estimate the Gaussian distribution parameters (μ, Σ) by the method of moments after each iteration of the Gibbs sampling. Moreover, we use fixed values for the hyperparameters, i.e., we set $\alpha = 50/K, \eta = 50/R, \beta = \gamma = \xi = 0.01, \lambda_1 = \lambda_2 = 0.5$ as Griffiths and Steyvers (2004) and Yin and Cui (2016) suggested.

According to the generative process of SCOMM, the joint probability distribution of the latent and observed variables can be factorized as follows:

$$
\begin{aligned}
& P\left(\boldsymbol{z}, \boldsymbol{a}, \boldsymbol{W_d}, \boldsymbol{M_d}, \boldsymbol{y}, \boldsymbol{l_d}, \boldsymbol{L_{M_d}} | \alpha, \lambda, \beta, \gamma, \eta, \xi, \boldsymbol{\mu}, \boldsymbol{\Sigma}\right) \\
& = P(\boldsymbol{z}|\alpha) P(\boldsymbol{y}|\lambda) P(\boldsymbol{W_d}|\boldsymbol{z}, \boldsymbol{y}, \beta) P(\boldsymbol{M_d}|\boldsymbol{z}, \boldsymbol{W_d}, \boldsymbol{y}, \gamma) P(\boldsymbol{a}|\eta) \\
& \quad \times P\left(\boldsymbol{L_{M_d}}|\boldsymbol{a}, \boldsymbol{\mu}, \boldsymbol{\Sigma}\right) P(\boldsymbol{l_d}|\boldsymbol{z}, \boldsymbol{a}, \xi) .
\end{aligned}
\tag{1}
$$

For Gibbs sampler, we need to derive the posterior probability for sampling latent topic z, latent switch y and latent area a for each mentioning activity. Hence, we first sample topic z according to the conditional probability $P(\boldsymbol{z_{(u,d)}} = k|\boldsymbol{z_{\neg(u,d)}}, \boldsymbol{a}, \boldsymbol{M_d}, \boldsymbol{W_d}, \boldsymbol{y}, \boldsymbol{L_{M_d}}, \boldsymbol{l_d}, \boldsymbol{u}\cdot)$, where $z_{\neg(u,d)}$ represents topic z assignments for all activities except the current one d of u. Due to space limitations, we omit the derivation details here. According to the Bayes chain rule and the joint probability distribution of all variables shown in Eq. (1), the sampling probability of a latent topic z is calculated as follows:

$$
\begin{aligned}
& P(\boldsymbol{z_{(u,d)}} = k|\boldsymbol{z_{\neg(u,d)}}, \boldsymbol{a}, \boldsymbol{M_d}, \boldsymbol{W_d}, \boldsymbol{y}, \boldsymbol{L_{M_d}}, \boldsymbol{l_d}, \boldsymbol{u}\cdot) \\
& \propto \frac{n^{D_u,k}_{\neg(u,d)} + \alpha}{\sum_{z' \in T}\left(n^{D_u,z'}_{\neg(u,d)} + \alpha\right)} \frac{n^{k,a,l_d}_{\neg(u,d)} + \xi}{\sum_{l' \in L_D}\left(n^{k,a,l'}_{\neg(u,d)} + \xi\right)} \\
& \quad \prod_{w \in W_d} \frac{n^{k,w}_{\neg(u,d)} + \beta}{\sum_{w' \in W}\left(n^{k,w'}_{\neg(u,d)} + \beta\right)} \\
& \quad \times \prod_{m \in M_d} \sum_{w \in W_d} \frac{n^{m,k,w}_{\neg(u,d)} + \gamma}{\sum_{m' \in M}\left(n^{m',k,w}_{\neg(u,d)} + \gamma\right)},
\end{aligned}
\tag{2}
$$

where $n^{D_u,k}$ is the number of activities assigned to topic k of u, n^{k,a,l_d} is the number of times that location l_d is sampled from the multinomial distribution over locations specific for topic k and area a, $n^{k,w}$ is the number of times word w is generated by topic z, $n^{m,k,w}$ is the number of times that word w and mentionee m co-occur in the same user activity under the topic k, $\neg(u, d)$ denotes that all the counts are calculated without taking account of the current activity d of u.

The, the coin indicator y can be sampled according to the following posterior probability.

$$
\begin{aligned}
& P\left(\boldsymbol{y_{(u,d,w)}} = 0|\boldsymbol{y_{\neg(u,d,w)}}, \boldsymbol{z}, \boldsymbol{a}, \boldsymbol{M_d}, \boldsymbol{W_d}, \boldsymbol{L_{M_d}}, \boldsymbol{l_d}, \boldsymbol{u}\cdot\right) \\
& \propto \frac{n^{y=0}_{\neg(u,d,w)} + \lambda}{\sum_{y' \in [0,1]}\left(n^{y=y'}_{\neg(u,d,w)} + \lambda\right)} \frac{n^{z,w,y=0}_{\neg(u,d,w)} + \beta}{\sum_{w' \in W}\left(n^{z,w',y=0}_{\neg(u,d,w)} + \beta\right)}
\end{aligned}
\tag{3}
$$

$$
\begin{aligned}
& P\left(\boldsymbol{y_{(u,d,w)}} = 1|\boldsymbol{y_{\neg(u,d,w)}}, \boldsymbol{z}, \boldsymbol{a}, \boldsymbol{M_d}, \boldsymbol{W_d}, \boldsymbol{L_{M_d}}, \boldsymbol{l_d}, \boldsymbol{u}\cdot\right) \\
& \propto \frac{n^{y=1}_{\neg(u,d,w)} + \lambda}{\sum_{y' \in [0,1]}\left(n^{y=y'}_{\neg(u,d,w)} + \lambda\right)} \frac{n^{b,w,y=1}_{\neg(u,d,w)} + \beta}{\sum_{w' \in W}\left(n^{b,w',y=1}_{\neg(u,d,w)} + \beta\right)},
\end{aligned}
\tag{4}
$$

where the number $n^{y=0}$ is a count of topic words (i.e., words are generated from the topic-word distribution) and $n^{y=1}$ is a count of background words (i.e., words are generated from the background word distribution), $n^{z,w,y=0}$ is the number of times that word w appears as a topic word and $n^{b,w,y=1}$ is the number of times word w occurs as a background word, and the number $n_{\neg(u,d,w)}$ denotes a quantity excluding the current word w in activity d of u.

At last, given the variable states except the latent area a, the sampling probability is calculated as follows:

$$
\begin{aligned}
& P\left(\boldsymbol{a_{(u,d)}} = x|\boldsymbol{a_{\neg(u,d)}}, \boldsymbol{z}, \boldsymbol{M_d}, \boldsymbol{W_d}, \boldsymbol{y}, \boldsymbol{L_{M_d}}, \boldsymbol{l_d}, \boldsymbol{u}\cdot\right) \\
& \propto \frac{n^{u,x}_{\neg(u,d)} + \eta}{\sum_{a' \in A}\left(n^{u,a'}_{\neg(u,d)} + \eta\right)} \frac{n^{z,x,l_d}_{\neg(u,d)} + \xi}{\sum_{l' \in L_D}\left(n^{z,x,l'}_{\neg(u,d)} + \xi\right)} \prod_{m \in M_d} P(l_m|\mu_a, \Sigma_a),
\end{aligned}
\tag{5}
$$

where $n^{u,x}$ is the number of times that area a has been sampled from u, and the number $n_{\neg(u,d)}$ denotes a quantity excluding the current instance. Moreover, after each sampling, we employ the method of moments to update the Gaussian distribution parameters (i.e., μ and Σ) according to the assigned area a for simplicity and speed. Specifically, parameters μ_a and Σ_a are updated as follows:

$$\mu_a = E(a) = \frac{1}{|S_a|}\sum_{m\in S_a} l_m \tag{6}$$

$$\Sigma_a = D(a) = \frac{1}{|S_a|-1}\sum_{m\in S_a}(l_m-\mu_a)(l_m-\mu_a)^T, \tag{7}$$

where S_a denotes the set of mentionees assigned with area a.

After a sufficient number of sampling iterations, we can make the following estimation of the model parameters with the collapsed Gibbs sampler by using the calculated approximate posteriors.

$$\theta^{u,z} = \frac{n^{u,z}+\alpha}{\sum_{z'\in T}\left(n^{u,z'}+\alpha\right)} \tag{8}$$

$$\phi^{u,a} = \frac{n^{u,a}+\eta}{\sum_{a'\in A}\left(n^{u,a'}+\eta\right)} \tag{9}$$

$$\vartheta^{z,w} = \frac{n^{z,w}+\beta}{\sum_{w'\in W}\left(n^{z,w'}+\beta\right)} \tag{10}$$

$$\psi^{z,w,m} = \frac{n^{z,w,m}+\gamma}{\sum_{m'\in M}\left(n^{z,w,m'}+\gamma\right)} \tag{11}$$

$$\pi^{z,a,l_d} = \frac{n^{z,a,l_d}+\xi}{\sum_{l'\in L_D}\left(n^{z,a,l'}+\xi\right)} \tag{12}$$

4. Mentionee recommendation

In this section, we show how to apply the learned SCOMM model (i.e., the parameter set $\hat{\Phi} = \left\{\hat{\theta}, \hat{\vartheta}, \hat{\psi}, \hat{\phi}, \hat{\pi}, \hat{\mu}, \hat{\Sigma}\right\}$) to the mentionee recommendation task.

Given a query $q = \left(u_q, l_q, W_q\right)$, i.e., mentioner u_q publishes a mentioning post q with words W_q at location l_q, the probability of m being mentioned can be calculated as follows:

$$P\left(m|u_q, l_q, W_q, \hat{\Phi}\right) = \sum_{z\in T} P\left(z|u_q, \hat{\Phi}\right) P\left(m|l_q, W_q, z, \hat{\Phi}\right), \tag{13}$$

where the component $P\left(m|l_q, W_q, z, \hat{\Phi}\right)$ is calculated according to the Bayes rules as follows:

$$\begin{aligned} P\left(m|l_q, W_q, z, \hat{\Phi}\right) &= \frac{P\left(m, l_q|W_q, z, \hat{\Phi}\right)}{\sum_{m'\in M} P\left(m', l_q|W_q, z, \hat{\Phi}\right)} \\ &\propto P\left(m, l_q|W_q, z, \hat{\Phi}\right) \\ &\propto \sum_{a\in A}\Bigg[P(a)\, P\left(l_m|a, \hat{\Phi}\right) P\left(l_q|z, a, \hat{\Phi}\right) \\ &\qquad \prod_{w\in W_q} P\left(w|W_q\right) P\left(m|W_q, z, \hat{\Phi}\right)\Bigg], \end{aligned} \tag{14}$$

where $P\left(w|W_q\right)$ denotes the weight of a word w in W_q and is estimated using Inverse Document Frequency (IDF) in this work. The prior probability of area a can be estimated as follows:

$$P(a) = \sum_{u\in U} P(a|u)\, P(u) = \sum_{u\in U} P(u)\, \hat{\phi}_{u,a} \tag{15}$$

$$P(u) = \frac{|D_u|+\varepsilon}{\sum_{u'\in U}\left(|D'_u|+\varepsilon\right)}, \tag{16}$$

where $|D_u|$ denotes the number of mentioning records generated by u and we use a Dirichlet prior parameter ε to play the role of a pseudocount to avoid overfitting as Yin and Cui (2016) suggested.

Based on the equations above, we reformulate Eq. (13) into Eq. (17). Then, we retrieve the top-k mentionees with the highest probabilities as the recommendations.

$$\begin{aligned} & P\left(m|u_q, l_q, W_q, z, \hat{\Phi}\right) \\ & \propto \sum_{z\in T}\Bigg[P\left(z|u_q, \hat{\Phi}\right)\sum_{a\in A}\Bigg[P(a)\, P\left(l_m|a, \hat{\Phi}\right) P\left(l_q|z, a, \hat{\Phi}\right) \\ & \qquad \prod_{w\in W_q} P\left(w|W_q\right) P\left(m|W_q, z, \hat{\Phi}\right)\Bigg]\Bigg] \\ & = \sum_{z\in T}\Bigg[\hat{\theta}_{z,u_q}\sum_{a\in A}\Bigg[P(a)\, \hat{\pi}_{l_q,z,a} P\left(l_m|\hat{\mu}_a, \hat{\Sigma}_a\right) \\ & \qquad \prod_{w\in W_q} P\left(w|W_q\right)\hat{\psi}_{m,z,w}\Bigg]\Bigg] \\ & = \sum_{z\in T}\sum_{a\in A}\Bigg[\hat{\theta}_{z,u_q} P(a)\, \hat{\pi}_{l_q,z,a} P\left(l_m|\hat{\mu}_a, \hat{\Sigma}_a\right)\prod_{w\in W_q} P\left(w|W_q\right)\hat{\psi}_{m,z,w}\Bigg] \end{aligned} \tag{17}$$

4.1. Efficient top-k recommendation

As discussed in Section 1, the time efficiency is essential to our task. To speed up the online recommendation, we separate the computation of the ranking score $S(q, m)$ of a mentionee m for a given query q into the online scoring part and the offline scoring part according to Eq. (17), inspired by Yin and Cui (2016), Yin et al. (2016).

$$\begin{aligned} S(q, m) &= \sum_{t\in(z,a)} O(q, t)\, F(t, m) \\ &= \|\boldsymbol{q}\|\,\|\boldsymbol{m}\|\cos\left(\Delta_{\boldsymbol{q},\boldsymbol{m}}\right) \propto \cos\left(\Delta_{\boldsymbol{q},\boldsymbol{m}}\right) \end{aligned} \tag{18}$$

$$O(q, t) = \hat{\theta}_{z,u_q}\hat{\pi}_{l_q,z,a}\prod_{w\in W_q} P\left(w|W_q\right)\hat{\psi}_{m,z,w} \tag{19}$$

$$F(t, m) = P(a)\, P\left(l_m|\hat{\mu}_a, \hat{\Sigma}_a\right) \tag{20}$$

where t denotes an attribute in $T \times A$ set, i.e., $t \in \{(z, a)\, | z \in T, a \in A\}$. $F(t, m)$ represents the score of a mentionee m with respect to the attribute t, which is computed offline. On the other hand, $O(q, t)$ represents the query preference of q on attribute t, which is computed online. We can see that both $F(q, m)$ and the main components of $O(q, m)$ (i.e., $\hat{\theta}_{z,u_q}$, $\hat{\pi}_{l_q,z,a}$ and $\hat{\psi}_{m,z,w}$) are computed offline. In this way, the query time is significantly reduced by using an offline pre-computing process of the scores.

In this task, the most straightforward way of making recommendations is to calculate the recommendation scores of all candidates according to Eq. (18). Doing so requires us to traverse all the attributes for each mentionee, which is highly time-consuming. To improve the online query performance, a more efficient online retrieval process is needed to prune the search space in both semantic and spatial dimensions. For our problem, the ranking scores of mentionees for a query are calculated as the inner product between the query and mentionee vectors as shown in Eq. (18). As all mentionee vectors have the same length in our model, it gives us a straightforward solution to resolve this Maximum Inner Product Search (MIPS) problem by transforming it into finding K-Nearest Neighbors (KNN) among mentionees for each query through some manipulations as discussed in Shrivastava and Li (2014). Then, the problem can be solved approximately with many nearest neighbor search approaches, such as tree-index, grid-index and hybrid-index based semantic-spatial retrieve algorithms (Koenigstein, Ram, & Shavitt, 2012; Zhang, Chan, & Tan, 2014) and hashing based KNN search algorithms (Neyshabur & Srebro, 2015; Shrivastava & Li, 2014). However, the topic and area distributions in our model have very high dimensions, which severely degrades the pruning efficiency (also known as the "curse of dimensionality") of most single-dimensional search algorithms such as Metric-tree (Koenigstein et al., 2012). Although several improved high-dimensional indexing schemes alleviated the search time consumption by leveraging hybrid geo-textual indices, they are not constructed to retrieve user behavioral activities with highly semantic relevances. Besides, many of these methods suffer great accuracy losses since they are fundamentally designed for approximate rather than accurate retrieval problems.

Recently, Yin et al. introduced a series of branch and bound algorithms to reduce the computational cost of finding the top-ranked items after corresponding latent-class models have been trained, such as Threshold-based Algorithm (TA) (Yin et al., 2013), Clustering-based Branch and Bound algorithm (CBB) (Yin et al., 2016) and Attribute Pruning algorithm (AP) (Yin & Cui, 2016). Specifically, TA is an extension of the traditional threshold-based algorithm which may save the computation for some items through dynamic threshold, but it still suffers from the curse of dimensionality since it has to frequently update the threshold. To cope with the high dimensionality, AP constraints the attribute space by filtering out the attributes with low query preference and item relevance, while CBB prunes the item space by selecting the buckets of items with the high upper-bound ranking scores on each attribute. Both of these two solutions hold the nice property of terminating early without accessing all items or attributes. However, even though AP only needs to scan a few attributes of each item, it still needs to traverse through all items. As for CBB, it has to scan all the attributes for each selected item to compute the full ranking scores.

In this work, we designed an Item-Attribute Pruning (IAP) algorithm to prune the search space of both the items and attributes simultaneously, inspired by CBB (Yin et al., 2016) and AP (Yin & Cui, 2016). The proposed algorithm is based on two observations: (1) the inner product of the two vectors $\boldsymbol{q}$ and $\boldsymbol{m}$ is directly proportional to their directions since the values of the norm $\|\boldsymbol{q}\|$ and norm $\|\boldsymbol{m}\|$ are two constants, as shown in Eq. (18); (2) only when a query q prefers an attribute t and the item m has a high value on t, the score $O(q,t)\,F(t,m)$ will contribute significantly to the final ranking score. Hence, we first cluster the mentionee vectors into G groups according to their directions using the spherical k-means algorithm (Dhillon & Modha, 2001). For each group g, an upper-bound vector $\boldsymbol{g}$ is computed as the maximum $F(t,m)$ in the group on each attribute t (i.e., $\max_{m\in g} F(t,m)$). Clearly, $S(q,g)$ represents the upper-bound score for all mentionees in g since $O(q,t)$ is nonnegative. Moreover, for each attribute $t_i (1 \le i \le K \times Q)$, we compute an ordered list of mentionees V_{t_i}, where V_{t_i} consists of k mentionees sorted by the highest $F(t_i, m)$ values. And for each mentionee m, we rank its attributes according to the values of $F(t_i, m)$. Note that, all of these calculations and sorting are preformed offline.

For an online query q, the algorithm aims to: (1) find the candidate mentionee groups with potentially highest scores without accessing all groups; (2) for each mentionee in a group, calculate the ranking score with scanning only a few significant attributes for the mentionee. The online procedure of the IAP algorithm is as follows. First, IAP sorts the mentionee groups according to the inner product $S(q,g)$ between query vector $\boldsymbol{q}$ and the upper-bound vector $\boldsymbol{g}$ of each group (Line 1). Second, IAP finds k candidate mentionees according to the query preference and mentionee weight on their attributes (Lines 2–12). More specifically, we pick the top N attributes that cover most of the query preferences (i.e., $\sum_{i=1}^{n} O(q,t_i) > 0.9\sum_{j=1}^{K\times Q} O(q,t_j)$) with the smallest N (Line 3). For each of the top N attributes t, we choose the top ranked mentionees from V_t as candidates (Lines 4–12). Then, IAP sequentially accesses the sorted groups of mentionees. We adopted a binary min-heap to implement the result list V to ensure that the top m' has the smallest ranking score in V. Hence, for an unscanned group g, if $S(q,m')$ is no less than its upper-bound score $S(q,g)$, the algorithm terminates early without needing to check other groups (Lines 15–16). Otherwise, we examine each mentionee m in g and check whether we can avoid traversing all attributes for m, inspired by the AP algorithm and the region pruning technology proposed in Zhao, Cong, Yuan, and Zhu (2015) (Lines 18–30). Specifically, suppose we have scanned attributes $\{t_1, \ldots, t_{i-1}\}$, the partial score for these traversed attributes S_p is

$$\sum_{j=1}^{i-1} O(q,t_j)\,F(t_j,m)\,.$$

When it comes to the attribute t_i, the upper bound score for m is

$$S_p + \sum_{j=i}^{K\times Q} O(q,t_j)\,F(t_j,m)\,.$$

As we traverse the attributes in descending order of $F(t,m)$, the value of $F(t,m)$ for the remaining attributes $\{t_{i+1}, t_{i+2}, \ldots, t_{K\times Q}\}$ should be less than the one for the attribute t_i. Therefore, the partial score for the remaining attributes is at most

$$\sum_{j=i}^{K\times Q} O(q,t_j)\,F(t_j,m)\,.$$

In this case, the upper bound score of $S(q,m)$ for all attributes is

$$S_p + \sum_{j=i}^{K\times Q} O(q,t_j)\,F(t_j,m)\,.$$

If this upper bound is smaller than the ranking score of the top mentionee m', there is no need to check the remaining attributes (Lines 24–26). Otherwise, we compute the full score of the mentionee to compare with m' (Lines 27–30).

5. Experiment

In this section, we detail an experimental study of the mentionee recommendation problem and evaluate the performance of our methods compared with other state-of-the-art methods. The goal of our experiments is to understand: (1) whether the spatial context information helps in improving the performance of mentionee recommendation; (2) what impact does the various factors have on mentionee recommendation; (3) how the parameters of our model affect the quality of recommendation; and (4) when compared with other methods, whether IAP can achieve faster online mentionee recommendation.

Algorithm 2: The Online Item-Attribute Pruning Algorithm

```
Input: mentionee set $M$ in $G$ groups, query $q$, ranked
       mentionee list $V_t$
Output: result list $V$ with top-$k$ largest scores
Sort the groups by $S(q, b)$;
Compute the $\sum_{i=1}^{K\times Q} O(q, t_i)$ and sort the attributes by $O(q, t)$;
Pick top $N$ attributes satisfying:
$N \leftarrow \min\left(\left\{n \mid \sum_{i=1}^{n} O(q, t_i) > 0.9 \sum_{j=1}^{K\times Q} O(q, t_j), K \times Q\right\}\right)$;
for each attribute $t \in N$ do
    for each $m \in V_t$ and $m \notin V$ do
        if $V.size() < k$ then
            $V.insert(\langle m, S(q, m)\rangle)$;
        else
            $m' \leftarrow V.top()$;
            if $S(q, m) > S(q, m')$ then
                $V.delete(m')$;
                $V.insert(< m, S(q, m) >)$;
for each group $g_i \in G$ do
    $m' \leftarrow V.top()$;
    if $S(q, m') \geq S(q, b_i)$ then
        break;
    else
        for each $m \in g_i$ and $m \notin V$ do
            $S_p \leftarrow 0$; $O_p \leftarrow 0$; $Skip \leftarrow false$;
            $m' \leftarrow V.top()$;
            while there exists attribute $t$ not examined for $m$ do
                $S_p \leftarrow S_p + O(q, t) F(t, m)$;
                $O_p \leftarrow O_p + O(q, t)$;
                if
                $S_p + \left(\sum_{i=1}^{K\times Q} O(q, t_i) - O_p\right) F(t, m) \leq S(q, m')$
                then
                    $Skip \leftarrow true$;
                    break;
            if $Skip = false$ then
                if $S(q, m) > S(q, m')$ then
                    $V.delete(m')$;
                    $V.insert(< m, S(q, m) >)$;
$V.reverse()$;
return $V$;
```

5.1. Datasets and settings

5.1.1. Datasets

The experiments were conducted on two real-world datasets crawled from Sina Weibo and Twitter.

- **Weibo Dataset.** To construct the experiment dataset, we collected Weibo user profiles along with the following-follow networks of 5 initial users (each of them holds about 1,000 followees) following the work described in Gong et al. (2015). We selected users who had published at least 3 Geo-tagged posts between January 1, 2014 and July 1, 2014. Through these steps, we had an initial dataset contains 1,701,286 user profiles and 26,994,838 Geo-tagged posts. Since we focus on users' mention behaviors, we constructed the mentioning network by selecting posts with at least one mentioning instance (i.e., "@MentioneeName"). Then, we located mentionees' home locations (in the form of geographical coordinates) using a user home location acquiring method proposed in Compton et al. (2014). At last, the Weibo dataset we constructed consisted of 594,187 Geo-tagged mentioning posts published by 147,621 mentioners and 251,054 mentionees with known home locations.
- **Twitter dataset.** Using same crawling strategy, we crawled an initial Twitter dataset that contained 35,219,791 Geo-tagged tweets published by 1,620,081 users between October 15, 2016 and March 15, 2017. The final Twitter dataset with the constructed mention network contained 553,145 Geo-tagged mentioning tweets published by 133,625 mentioners in 807,330 mention-mentioned relations, as well as 187,843 mentionees with known home locations.

In the raw datasets, most of users' home locations were not publicly available or too coarse-grained to be identified as a specific location. To assign a home location to each mentionee, we adopted a simple and effective user home location acquiring approach used in many works (Compton et al., 2014; Jurgens, 2013; McGee et al., 2013). Specifically, for a mentionee, we first calculated the $l1$-multivariate median (Vardi & Zhang, 2000) of all her tagged locations as her home location. The reason why we opted to use a median location rather than a mean is that, the geometric median is robust to location outliers, such as when an individual published Geo-tagged posts from an atypical location far from the normal concentration of locations. Then, we filtered out bots and spammers with unnatural movement trajectories. We only keep users with at least three GPS-tagged tweets that occur within a 15 km geographic radius, and filtered out users who have traveled in excess of 1,000 km/h according to the tagged timestamps and locations. Furthermore, for those who specified a detailed home location (e.g., GPS coordinates or detailed location description), the self-reported locations were used as their ground truth home locations by searching through the geographical database GeoNames.[5] and the AMAP Geocoder API.[6] Table 2 lists the descriptive statistics of the two datasets.

Note that, for a mentioning activity $d = (u, W_d, l_d, M_d)$, since the norm of M_d may be greater than 1, we denote each (u, W_d, l_d, m), $m \in M_d$ as a training/test case. Hence, there may be two or more cases with the same mentioner, words and tagged location, but different mentionees. In other words, the number of mentioning cases in the two datasets was decided by the number of mention relations rather than the number of posts. Furthermore, we adopted the time-dependent segmentation method to split the datasets into training and testing sets following the work in Gong et al. (2015). Specifically, we ranked all the mentioning records according to the time they were posted, and then split the records with an 80/20 proportion.

5.1.2. Evaluation metrics

For each test case (u, W_d, l_d, m) in the test set, we formed a recommendation list V_k by selecting the k mentionees with the top-ranked scores. For a ground-truth mentionee m, if $m \in V_k$, we considered it as a hit; otherwise, we considered it as a miss. To evaluate the recommendation effectiveness, we used the well known **Precision**, **Recall**, and **F1-Score** measures for the highest ranked results. We also used the Mean Reciprocal Rank (**MRR**) metric to measure the rank of the recommended results. At last, we adopted the measurement **Accuracy@**k to evaluate whether mentionees were correctly recommended from the top k results as follows:

$$Accuracy@k = \frac{N_{hits@k}}{N_{testcase}}$$

[5] http://www.geonames.org/.

[6] http://lbs.amap.com/api/javascript-api/guide/map-data/geocoding.

Table 2
Statistics of constructed datasets.

	Weibo	Twitter
# of posts	594,187	553,145
# of total users	309,416	261,949
# of mentioners	147,621	133,625
# of mentionees	251,054	187,843
# of mention-mentioned relations	1,300,259	807,330
# of Avg. mentionees per mentioner	8.81	6.04
# of Avg. mentions per post	2.19	1.46

where we use $N_{hits@k}$ to denote the number of hit times for a given k in the test set, and $N_{testcase}$ represents the number of all test cases.

5.2. Comparative approaches

Recommendation Effectiveness. For comparison with our proposed SCOMM, we evaluated the following methods in terms of recommendation effectiveness.

- HIStory based mentionee recommendation (**HIS**): HIS is a simple heuristic method that recommends the mentionee who is frequently mentioned by the mentioner. The candidates in the list are ranked in descending order of the frequency. If two or more mentionees share the same frequency, the latest mentioned one is selected.
- Context-aware At Recommendation (**CAR**): CAR, proposed in Tang et al. (2015), is a context-aware approach that aims to recommend target audiences for promotion-oriented messages. CAR measures the relevance among mentioners, mentionees and posts by extracting content, social, location and time-based features. Then, top-k results are produced based on learning-to-rank techniques. In our work, we implement this method using the same features in Tang et al. (2015). Hence, we crawled the historical posts (not necessarily with a location tag) of each mentionee and the interaction network between users (i.e., the replying and retweeting histories of users) to enrich our dataset.
- "At" User–User Topic Translation Model (**A-UUTTM**): A-UUTTM is a translation based model proposed by Gong et al. in Gong et al. (2015). A-UUTTM takes into account both the current content and the histories of mentionees, and is the state-of-the-art approach for mentionee recommendation. Since the histories of mentionees are important to A-UUTTM, we extracted the latest 4 posts published by each mentionee before she/he was mentioned. Moreover, we adopted the same hyperparameters used in A-UUTTM.
- SCOMM with no content information(**SCOMM-NT**): SCOMM-NT is a variant of our proposed SCOMM, which considers only the spatial factors, i.e., the words of the mentioning activities were ignored. In SCOMM-NT, a mentioner chooses a series of mentionees according to a topic-mentionee distribution.
- SCOMM with no spatial factors (**SCOMM-NS**): SCOMM-NS is another variant of our proposed SCOMM, which ignores the impact of mentioners' movement patterns and the mentionee clustering areas. It takes only textual content information into account and produces recommendations based on the basic topic-translation model.
- SCOMM with no mentionee clustering area modeling (**SCOMM-NA**): SCOMM-NA is a variant model of our proposed SCOMM with the mentionee clustering area modeling procedure removed. In SCOMM-NA, a topic z is responsible for generating words W_d, mentionee m, and location l_d. And mentionees are generated according to the topical translation process.
- SCOMM with no mentioner movement pattern modeling (**SCOMM-NP**): SCOMM-NP is a variant model of our proposed SCOMM without considering mentioners' movement patterns, i.e., the location tag l_d of mentioning activity d is ignored in the model.

Recommendation Efficiency. To demonstrate the efficiency of our proposed IAP algorithm, we compared our approach with the following four algorithms.

- Linear Scan algorithm (**LS**): LS is a baseline method that computes the ranking score after scanning all attributes for each mentionee and selects top-k ones with the highest scores at query time.
- Threshold algorithm (**TA**): TA (Yin et al., 2013) is an extension of the traditional threshold-based algorithm, which pre-computes a sorted list for each latent attribute and maintains a priority list of all the sorted lists that controls which list to access next.
- Clustering-based Branch and Bound algorithm (**CBB**): CBB (Yin et al., 2016) is an item space pruning algorithm designed for fast retrieval on high-dimensional data. Specifically, given a query, CBB first obtains an ordered list of all item buckets according to the ranking scores of pre-computed upper-bound vector of each bucket. Then, CBB prunes the item space by selecting the item buckets with the high upper-bound ranking scores on each attribute. For each item in the selected buckets, CBB calculates its final ranking score by sequentially accessing all attributes.
- Attribute Pruning algorithm (**AP**): AP (Yin & Cui, 2016) is an attribute space pruning algorithm designed for high-dimensional retrieval. Given a query, AP first selects k potential items that are good for recommendation according to the pre-sorted item lists for each attribute and the pre-sorted attribute lists of each item. Then, AP traverses the whole item spaces to check whether there are better choices than those in the potential list. In this step, AP prunes the attribute searching space for each item by filtering out the attributes with low query preference and item relevance.

5.3. Results and discussion

5.3.1. Recommendation effectiveness

In this section, we analyze the results of all the methods mentioned in Section 4.2 as well as our proposed SCOMM with well-tuned parameters in terms of recommendation effectiveness.

Recommendations on Weibo. Table 3 shows the performance of all methods on our Weibo and Twitter datasets in metrics Precision, Recall, F1-Score, MRR, Accuracy@3 and Accuracy@5. The experimental results on the Weibo dataset reported in 3a suggest that SCOMM outperformed all other methods significantly and consistently in all measures. It achieved: 0.663 in Precision, 0.65 in Recall, and 0.657 in F1-score. Compared with the state-of-the-art approach A-UUTTM, SCOMM performed better by 13.9% in Precision, 16.5% in Recall and 15.3% in F1-Score. The results of Accuracy@3 and Accuracy@5 showed that 68.2% of ground truth mentionees were correctly found in the top 3 and 71.3% of mentionees were correctly recommended in the top 5. The MRR result of SCOMM was also the best among the rest, which demonstrated that SCOMM was capable of producing not only more accurate recommendations but also better ranking of the results. Fig. 4a shows the Precision, Recall and F1-Score of all methods with different numbers of recommended users. We list metric values with k varying from 1 to 7 since the values do not change significantly when $k > 7$. In Fig. 5a, we report the top-k recommendation accuracies of SCOMM and the comparative methods. We show

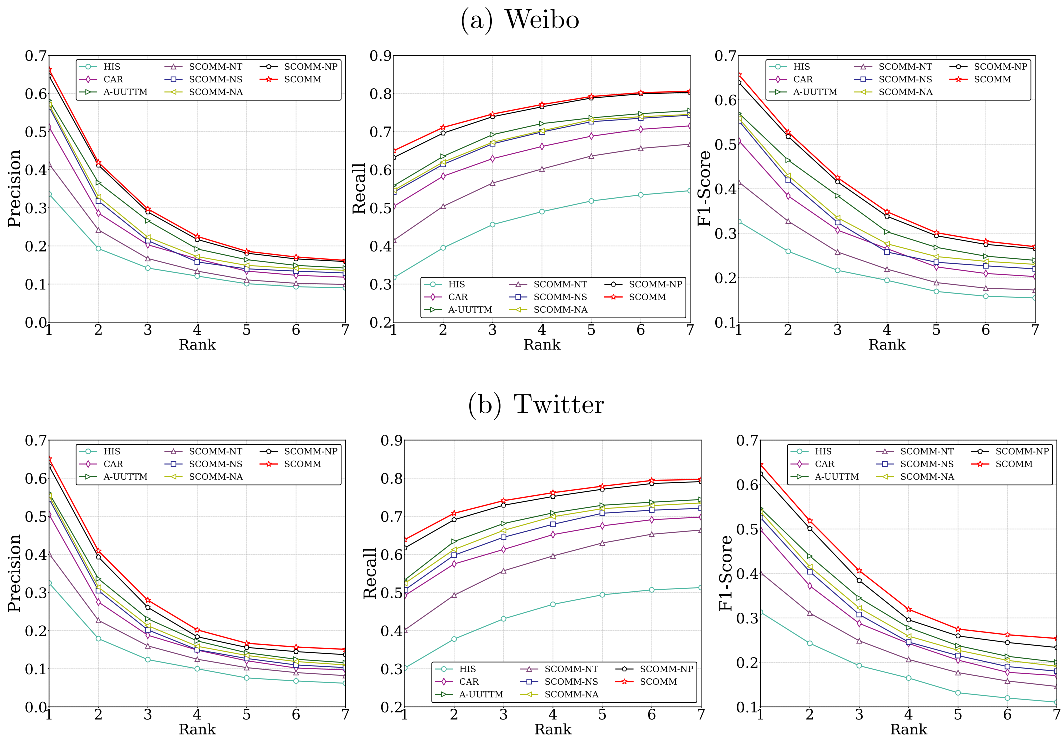


Fig. 4. Precision, Recall and F1-Score with different values of *k*.

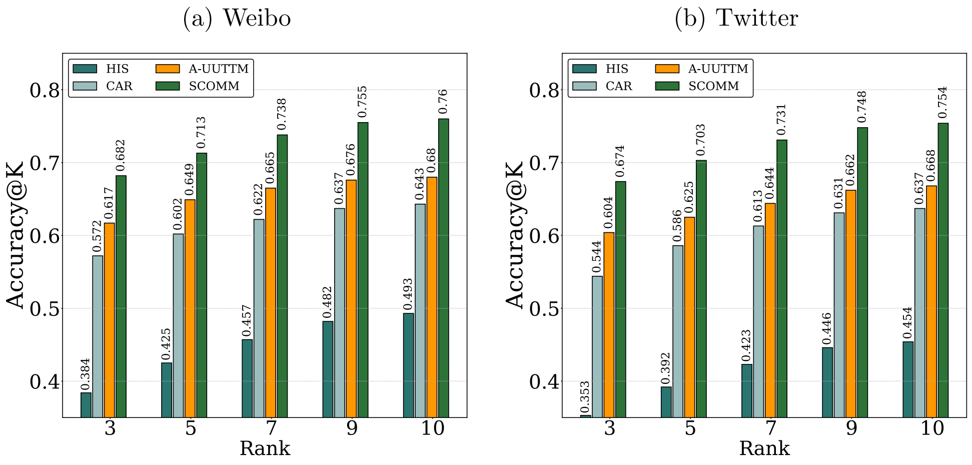


Fig. 5. Top-*k* recommendation accuracy.

only the performance where *k* is less than 10, because larger values of k are not practically useful for this recommendation task. The performance of SCOMM clearly demonstrates the advantages of jointly exploiting the semantic and spatial patterns of users' mentioning activities.

From the results, several additional observations can be made:

1. HIS, which is used by many real-world services, underperformed all other methods on all metrics.
2. Compared to CAR, A-UUTTM and SCOMM, SCOMM-NT performed poorly, showing the necessity of exploiting the semantic patterns of users' mentioning records to capture their interests. In SCOMM-NT, users' mentioning preference was inferred purely from the spatial context information of user mentioning activities, while ignoring the impact of the content.
3. Both SCOMM and A-UUTTM performed much better than CAR, which validated the advantages of a well-designed probabilistic generative model in the representation and generalization for recommendation over the general feature-based learning-to-rank method.
4. SCOMM achieved significantly better results than A-UUTTM, although A-UUTTM models user mention behavior based on the content of both mentioners and mentionees, which showed the benefits of exploiting the spatial context information of users' mentioning activities. Compared with A-UUTTM, SCOMM incorporates the effects of mentionees' geographical distribution based on their home locations, and the influence of mentioners' movement patterns based on the joint effect of their topical interest and the geographical distribution of their mentionees. Both CAR and A-UUTTM ignored the impact of these two spatial context factors on users' mention behaviors.

Table 3
The performance on Precision, Recall, F1-Score, MRR, Acc@3 and Acc@5 metrics.

Methods	Precision	Recall	F1-Score	MRR	Acc@3	Acc@5
(a) Weibo						
HIS	0.336	0.317	0.326	0.588	0.384	0.425
CAR	0.512	0.504	0.508	0.572	0.572	0.602
A-UUTTM	0.582	0.558	0.570	0.603	0.617	0.649
SCOMM-NT	0.415	0.381	0.397	0.422	0.451	0.475
SCOMM-NS	0.565	0.541	0.553	0.566	0.588	0.615
SCOMM-NA	0.571	0.546	0.558	0.577	0.606	0.632
SCOMM-NP	0.646	0.631	0.638	0.652	0.662	0.688
SCOMM	**0.663**	**0.650**	**0.657**	**0.678**	**0.682**	**0.713**
(b) Twitter						
HIS	0.325	0.302	0.313	0.575	0.353	0.392
CAR	0.505	0.492	0.498	0.552	0.544	0.586
A-UUTTM	0.561	0.533	0.547	0.581	0.604	0.625
SCOMM-NT	0.402	0.368	0.384	0.414	0.441	0.467
SCOMM-NS	0.546	0.507	0.526	0.553	0.554	0.579
SCOMM-NA	0.554	0.524	0.539	0.569	0.575	0.612
SCOMM-NP	0.632	0.617	0.624	0.638	0.653	0.681
SCOMM	**0.651**	**0.639**	**0.645**	**0.667**	**0.674**	**0.703**

Table 4
Top-k recommendation accuracy of variant methods and our approach.

Datasets	Methods	Accuracy@k				
		$k = 3$	$k = 5$	$k = 7$	$k = 9$	$k = 10$
Weibo	SCOMM-NT	0.451	0.475	0.494	0.509	0.514
	SCOMM-NS	0.588	0.615	0.630	0.644	0.649
	SCOMM-NA	0.606	0.632	0.646	0.658	0.662
	SCOMM-NP	0.662	0.688	0.711	0.723	0.729
	SCOMM	**0.682**	**0.713**	**0.738**	**0.755**	**0.760**
Twitter	SCOMM-NT	0.441	0.467	0.489	0.505	0.512
	SCOMM-NS	0.554	0.579	0.595	0.606	0.611
	SCOMM-NA	0.575	0.612	0.627	0.641	0.647
	SCOMM-NP	0.653	0.681	0.705	0.717	0.721
	SCOMM	**0.674**	**0.703**	**0.731**	**0.748**	**0.754**

Table 5
Recommendation Accuracy@5 with various K and Q.

Q	K					
	$K = 50$	$K = 60$	$K = 70$	$K = 80$	$K = 90$	$K = 100$
$Q = 40$	0.636	0.654	0.667	0.673	0.674	0.675
$Q = 50$	0.656	0.674	0.686	0.693	0.693	0.694
$Q = 60$	0.668	0.686	0.698	0.704	0.704	0.706
$Q = 70$	0.675	0.694	0.707	**0.713**	0.713	0.714
$Q = 80$	0.676	0.695	0.707	0.713	0.713	0.714
$Q = 90$	0.676	0.696	0.708	0.713	0.714	0.716

5. The content and spatial context-aware methods performed significantly better than single-factor aware latent class models. For example, the accuracy of SCOMM is 0.76 when $k = 10$, which makes a relative improvement of 47.9% compared to the spatial information based SCOMM-NT, and a 11.8% improvement compared to the content based A-UUTTM. This observation demonstrates the advantages of the model which considers both content and spatial factors synthetically in capturing users' mention behavioral preferences.

Through comparing the results of SCOMM and its variant models, we can further evaluate the impact on effectiveness of semantic patterns (S), mentioner movement patterns (M) and mentionee geographical distribution (D) factors on recommendation performance. We summarize the Accuracy@k performance of SCOMM-NT, SCOMM-NS, SCOMM-NA, SCOMM-NP and SCOMM in Table 4, where the larger a value is, the less important the missing factor of the model is. From the results, we first observe that SCOMM consistently outperformed the four variants, indicating that our proposed SCOMM benefited from synthetically considering the three factors. Second, we observe that the contribution of each factor to the improvement over the recommendation accuracy is different. According to Table 3a and 4, the order of importance of the three factors is $S \succ D \succ M$. First, the semantic patterns play a dominant role in improving mentionee recommendation performance. The accuracy of SCOMM-NP, which considers only the two spatial factors, is over 20% lower than that of SCOMM, which uses the same spatial factors but also takes into account the content of the posts. Our observation is consistent with previous studies (Chen et al., 2012; Tang et al., 2015; Zhou et al., 2015). Second, the geographical clustering areas of mentionees, although not as important as content, are still very helpful in improving model accuracy. Compared with SCOMM-NS, SCOMM-NP incorporated the mentionee home location modeling procedure, which leads to a 17.1% accuracy improvement when $k = 10$. Third, the performance of the methods that considered mentioners' movement patterns is better than those that did not. For example, SCOMM achieved a 3.6% improvement in *Accuracy*@5 compared with SCOMM-NP.

Recommendations on Twitter. From Table 3b and Fig. 4b, we can see that experiment on the Twitter dataset shows the similar result that SCOMM consistently outperformed all competitors. At the same time, compared to the experimental results on the Weibo dataset, two differences can also be observed. First, all methods achieved lower performance on the Twitter datset. This may be caused by the fact that we had fewer mentioning records in the Twitter dataset, which led to a less accurate capturing of users' semantic and spatial preference from the model. Second, the performance gaps between methods that considered the geographical information and methods that did not were slightly larger on the Twitter dataset. For example, on the Weibo dataset, the accuracy gap between SCOMM and A-UUTTM was 6.4% when $k = 5$, and 8% when $k = 10$; while, on the Twitter dataset, the accuracy gap was increased to 7.8% when $k = 5$ and 8.6% when $k = 10$. We believe this may be because the location data in the Twitter dataset were less sparse than in the Weibo dataset.

5.3.2. Parameters sensitivity analysis

We also conducted a series of experiments to study the impact of tuning model parameters on the performance, i.e., the number of topics K and the number of areas Q. We only show the experimental results on the Weibo dataset, since the results on the Twitter dataset are similar. Specifically, we show how the Accuracy@5 performance of our proposed SCOMM changes as K varies from 50 to 100 and Q from 40 to 120. As for the hyperparameters, we fixed them at $\alpha = 50/K$, $\eta = 50/R$, $\beta = \gamma = \xi = 0.01$, $\lambda_1 = \lambda_2 = 0.5$ as Griffiths and Steyvers (2004) and Yin et al. (2016) suggested. Table 5 reports the performance of SCOMM on varying numbers of topics and areas. We observed that the recommendation accuracy of SCOMM increased quickly first with the increasing values of K and Q. Then the increase slowed down when K and Q exceed certain thresholds (i.e., $K = 80$ and $Q = 70$). The reason is that the numbers of the topics and areas reflect the model complexity. When K and R were too small, the model has limited ability to describe the data. On the other hand, when the values of K and R reached certain thresholds, the model was complex enough to handle the data and a further increasing of the numbers did not help much.

Moreover, to evaluate the best performance and the performance changes under larger parameter values of our model, we conducted another experiment by varying K and Q on a large scale. Table 6 shows the Accuracy@5 values when $K > 190$ and $Q > 210$. From the results, we can see that the proposed SCOMM achieved its best result when $K = 210$ and $Q = 230$. However, the accuracy dropped significantly when K and Q are increased

Table 6
Recommendation Accuracy@5 with large K and Q.

Q	K				
	$K = 190$	$K = 200$	$K = 210$	$K = 220$	$K = 230$
$Q = 210$	0.724	0.724	0.724	0.717	0.706
$Q = 220$	0.724	0.724	0.725	0.717	0.709
$Q = 230$	0.724	0.724	**0.725**	0.718	0.709
$Q = 240$	0.723	0.722	0.719	0.711	0.703
$Q = 250$	0.722	0.721	0.718	0.708	0.699

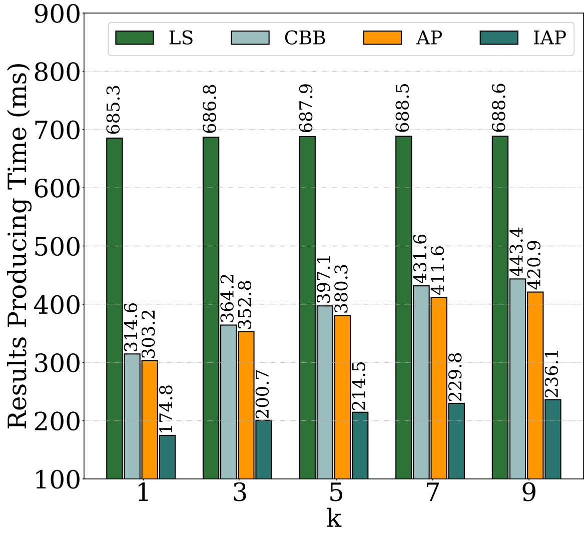


Fig. 6. Recommendation efficiency on Weibo dataset.

further. A possible reason is that when the values of K and Q were too large, the data sparsity problem became more significant when estimating the topic and area-specific probability matrix (e.g., ψ), which led to overfitting and made the learned parameters unreliable. Considering the trade-off between model effectiveness and model efficiency, the performance reported in Section 5.3.1 is achieved with the parameter settings $K = 80$ and $Q = 70$ according to Table 5.

5.3.3. Recommendation efficiency

In this section, we evaluate the online recommendation efficiency of the proposed IAP on the Weibo dataset by comparing IAP with the algorithms described in Section 5.1.2. All the online algorithms were implemented in Java (JDK 1.7) and ran on a Windows Server 2008 R2 system with an Intel Xeon E5 processor (2.4 GHz) and 160G RAM.

We show the algorithm performance with 5,600 latent dimensions ($K = 80$ and $R = 70$) and with k set to 1, 3, 5, 8 and 10. We tested all queries created in the test set. The number of groups G used in the IAP algorithm was set to 420 after trying different numbers. Fig. 6 presents the average top-k result producing time of the LS, CBB, AP and IAP algorithms. Note that we did not list the result producing time of TA in Fig. 6, because the efficiency performance of TA was way worse than other four algorithms. In fact, TA took 5,719.6 ms to produce the top-1 result and spent nearly 7,000 ms to find the top-5 recommendations. The reason TA performed so poorly is that it has to frequently update the threshold and to maintain the dynamic priority queue of the sorted lists, which is very time-consuming when the dimensionality is high.

From the results, we can first observe that our purposed IAP outperformed all competitors significantly and delivered the best results robustly. For example, IAP found the right top-5 results from over 250,000 mentionees in 229.8 ms, which was over 3 times faster than the brute-force algorithm LS, 1.8 times faster than the attribute pruning algorithm AP, and about 1.9 times faster than the item pruning algorithm CBB. Although the performance of IAP deteriorated as the number of recommendations k increased, it was still much faster than all other algorithms even when $k = 10$. This result demonstrates the advantages of pruning the search space of both the items and attributes simultaneously. In our experiment, IAP only needed to examine about 75,941 mentionees on average (about 30.2% of all mentionees) when $k = 5$. And for each of these mentionees, IAP only needed to scan 469 attributes on average (about 8.4% of all attributes) for $k = 5$. After all, the number of mentionees that need to be scanned and computed for the full scored by accessing all the attributes is 9,354, which was only 3.7 percent of all mentionees. From the above, we can see that our proposed IAP is more efficient than LS, TA, CBB and AP, especially when the data dimensionality is very high.

6. Conclusion

In this work, we aim to tackle the essential problem of mentionee recommendation, i.e., how to find mentionees who are most likely to be noticed by a common user for knowing a post. Specifically, we investigate the problem by exploring users' online mention behaviors. After analyzing two large real-world social media datasets, we found that users' mention behaviors are influenced by not only the semantic but also the spatial context factors of their mentioning activities. More specially, we observed that the home locations of mentionees for a mentioner tend to be within a specific geographic area, and the mentioners are more likely to select mentionees who live close to their standing locations. Based on that, we proposed a joint latent-class probabilistic model SCOMM to simulate the process of generating users' mentioning activities. By introducing two types of latent topics to generate the semantic and geographical attributes of users' mentioning activities, SCOMM simultaneously learns and models the semantic patterns of mentioners, the geographical clustering areas of mentionees, and their joint effects on mentioners' movement patterns. Based on SCOMM, we designed an Item-Attribute Pruning algorithm to overcome the curse of dimensionality and to facilitate online top-k query performance by pruning the item and attribute space simultaneously. Extensive experiments were conducted to evaluate the performance of our proposed solution on two real-world datasets crawled from different social media sites. The experimental results showed the benefits of the spatial context information of users' mentioning activities in improving mentionee recommendation while demonstrating the superiority of our approach over other state-of-the-art methods. Besides, we carried out an ablation study to evaluate the impact of different factors in SCOMM, and found that: (1) the semantic patterns of mentioners play a dominant role in improving mentionee recommendation; (2) the geographical clustering areas of mentionees are very helpful in improving model performance; (3) models which considered mentioners' movement patterns performed better than those that did not.

Acknowledgment

This work is supported by the National Natural Science Foundation of China (No. 61502350).

References

Alam, M. H., Ryu, W. J., & Lee, S. (2017). Hashtag-based topic evolution in social media. *World Wide Web*, 1–23.

Bi, B., & Cho, J. (2016). Modeling a retweet network via an adaptive bayesian approach. In *Proceedings of the 25th international conference on world wide web* (pp. 459–469). International World Wide Web Conferences Steering Committee.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, *3*(Jan), 993–1022.

Bobadilla, J., Ortega, F., Hernando, A., & Gutiérrez, A. (2013). Recommender systems survey. *Knowledge-Based Systems*, *46*, 109–132.

Brown, C., Nicosia, V., Scellato, S., Noulas, A., & Mascolo, C. (2012). The importance of being placefriends: discovering location-focused online communities. In *Proceedings of the 2012 ACM workshop on workshop on online social networks* (pp. 31–36). ACM.

Chen, K., Chen, T., Zheng, G., Jin, O., Yao, E., & Yu, Y. (2012). Collaborative personalized tweet recommendation. In *Proceedings of the 35th international ACM SIGIR conference on research and development in information retrieval* (pp. 661–670). ACM.

Cheng, Z., Caverlee, J., Barthwal, H., & Bachani, V. (2014). Who is the barbecue king of texas? A geo-spatial approach to finding local experts on twitter. In *Proceedings of the 37th international ACM SIGIR conference on research & development in information retrieval* (pp. 335–344). ACM.

Compton, R., Jurgens, D., & Allen, D. (2014). Geotagging one hundred million twitter accounts with total variation minimization. In *Big data, 2014 IEEE International conference on* (pp. 393–401). IEEE.

Dhillon, I. S., & Modha, D. S. (2001). Concept decompositions for large sparse text data using clustering. *Machine Learning*, *42*(1), 143–175.

Ding, Z., Qiu, X., Zhang, Q., & Huang, X. (2013). Learning topical translation model for microblog hashtag suggestion. In *IJCAI* (pp. 2078–2084).

Fang, Q., Xu, C., Hossain, M. S., & Muhammad, G. (2016). Stcaplrs: A spatial-temporal context-aware personalized location recommendation system. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *7*(4), 59.

Ge, H., Caverlee, J., & Lu, H. (2016). TAPER: A contextual Tensor-Based Approach for Personalized Expert Recommendation. In *RecSys* (pp. 261–268).

Gong, Y., Zhang, Q., Sun, X., & Huang, X. (2015). Who will you@? In *Proceedings of the 24th ACM international on conference on information and knowledge management* (pp. 533–542). ACM.

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences*, *101*(Suppl. 1), 5228–5235.

Gupta, P., Goel, A., Lin, J., Sharma, A., Wang, D., & Zadeh, R. (2013). Wtf: The who to follow service at twitter. In *Proceedings of the 22nd international conference on world wide Web* (pp. 505–514). ACM.

Hannon, J., Bennett, M., & Smyth, B. (2010). Recommending twitter users to follow using content and collaborative filtering approaches. In *Proceedings of the fourth ACM conference on recommender systems* (pp. 199–206). ACM.

Hofmann, T. (1999). Probabilistic latent semantic analysis. In *Proceedings of the fifteenth conference on uncertainty in artificial intelligence* (pp. 289–296). Morgan Kaufmann Publishers Inc.

Hu, B., & Ester, M. (2013). Spatial topic modeling in online social media for location recommendation. In *Proceedings of the 7th ACM conference on recommender systems* (pp. 25–32). ACM.

Huang, W., Kataria, S., Caragea, C., Mitra, P., Giles, C. L., & Rokach, L. (2012). Recommending citations: translating papers into references. In *Proceedings of the 21st ACM international conference on information and knowledge management* (pp. 1910–1914). ACM.

Jiang, B., Sha, Y., & Wang, L. (2015). Predicting user mention behavior in social networks. In *Natural language processing and chinese computing - 4th CCF conference, Proceedings* (pp.146–158). https://doi.org/10.1007/978-3-319-25207-0_13.

Jurgens, D. (2013). That's what friends are for: Inferring location in online social media platforms based on social relationships. *ICWSM*, *13*(13), 273–282.

Koenigstein, N., Ram, P., & Shavitt, Y. (2012). Efficient retrieval of recommendations in a matrix factorization framework. In *Proceedings of the 21st ACM international conference on information and knowledge management* (pp. 535–544). ACM.

Kogan, M., Palen, L., & Anderson, K. M. (2015). Think local, retweet global: Retweeting by the geographically-vulnerable during Hurricane Sandy. In *Proceedings of the 18th ACM conference on computer supported cooperative work & social computing* (pp. 981–993). ACM.

Kurashima, T., Iwata, T., Irie, G., & Fujimura, K. (2010). Travel route recommendation using geotags in photo sharing sites. In *Proceedings of the 19th ACM international conference on information and knowledge management* (pp. 579–588). ACM.

Lee, K., Mahmud, J., Chen, J., Zhou, M., & Nichols, J. (2014). Who will retweet this? Automatically identifying and engaging strangers on twitter to spread information. In *Proceedings of the 19th international conference on intelligent user interfaces* (pp. 247–256). ACM.

Li, Q., Song, D., Liao, L., & Liu, L. (2015). Personalized mention probabilistic ranking–recommendation on mention behavior of heterogeneous social network. In *International conference on web-age information management* (pp. 41–52). Springer.

Li, R., Wang, S., Deng, H., Wang, R., & Chang, K. C. C. (2012). Towards social user profiling: unified and discriminative influence model for inferring home locations. In *Proceedings of the 18th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1023–1031). ACM.

Lian, D., Xie, X., Zheng, V. W., Yuan, N. J., Zhang, F., & Chen, E. (2015). Cepr: a collaborative exploration and periodically returning model for location prediction. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *6*(1), 8.

Lim, K. H., Chan, J., Leckie, C., & Karunasekera, S. (2015). Detecting location-centric communities using social-spatial links with temporal constraints. In *European Conference on information retrieval* (pp. 489–494). Springer.

Liu, B., Fu, Y., Yao, Z., & Xiong, H. (2013). Learning geographical preferences for point-of-interest recommendation. In *Proceedings of the 19th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1043–1051). ACM.

McGee, J., Caverlee, J. A., & Cheng, Z. (2011). A geographic study of tie strength in social media. In *Proceedings of the 20th ACM international conference on information and knowledge management* (pp. 2333–2336). ACM.

McGee, J., Caverlee, J., & Cheng, Z. (2013). Location prediction in social media based on tie strength. In *Proceedings of the 22nd ACM international conference on information & knowledge management* (pp. 459–468). ACM.

Michelson, M., & Macskassy, S. A. (2010). Discovering users' topics of interest on twitter: a first look. In *Proceedings of the fourth workshop on analytics for noisy unstructured text data* (pp. 73–80). ACM.

Neyshabur, B., & Srebro, N. (2015). On symmetric and asymmetric lshs for inner product search. In *Proceedings of the 32nd international conference on international conference on machine learning* (Vol. 37) (pp. 1926–1934). JMLR. org.

Qiu, M., Zhu, F., & Jiang, J. (2013). It is not just what we say, but how we say them: Lda-based behavior-topic model. In *Proceedings of the 2013 SIAM international conference on data mining* (pp. 794–802). SIAM.

Scellato, S., Noulas, A., Lambiotte, R., & Mascolo, C. (2011). Socio-spatial properties of online location-based social networks. *ICWSM*, *11*, 329–336.

Shen, D., Ding, Z., Qiao, F., Cheng, J., & Wang, H. (2016). Finding the optimal users to mention in the appropriate time on twitter. In *International conference on knowledge science, engineering and management* (pp. 289–301). Springer.

Shrivastava, A., & Li, P. (2014). Asymmetric LSH (ALSH) for sublinear time maximum inner product search (MIPS). In *Advances in neural information processing systems* (pp. 2321–2329).

Takhteyev, Y., Gruzd, A., & Wellman, B. (2012). Geography of twitter networks. *Social Networks*, *34*(1), 73–81.

Tang, L., Ni, Z., Xiong, H., & Zhu, H. (2015). Locating targets through mention in Twitter. *World Wide Web*, *18*(4), 1019–1049.

Vardi, Y., & Zhang, C. H. (2000). The multivariate L1-median and associated data depth. *Proceedings of the National Academy of Sciences*, *97*(4), 1423–1426.

Wang, B., Wang, C., Bu, J., Chen, C., Zhang, W. V., Cai, D., et al. (2013). Whom to mention: expand the diffusion of tweets by@ recommendation on micro-blogging systems. In *Proceedings of the 22nd international conference on world wide web* (pp. 1331–1340). ACM.

Wang, W., Yin, H., Chen, L., Sun, Y., Sadiq, S., & Zhou, X. (2015). Geo-sage: A geographical sparse additive generative model for spatial item recommendation. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1255–1264). ACM.

Wang, Y., Yuan, N. J., Lian, D., Xu, L., Xie, X., Chen, E., et al. (2015). Regularity and conformity: location prediction using heterogeneous mobility data. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1275–1284). ACM.

Weng, J., Lim, E. P., Jiang, J., & He, Q. (2010). Twitterrank: finding topic-sensitive influential twitterers. In *Proceedings of the third ACM international conference on web search and data mining* (pp. 261–270). ACM.

Xu, Z., Zhang, Y., Wu, Y., & Yang, Q. (2012). Modeling user posting behavior on social media. In *Proceedings of the 35th international ACM SIGIR conference on research and development in information retrieval* (pp. 545–554). ACM.

Ye, M., Shou, D., Lee, W. C., Yin, P., & Janowicz, K. (2011). On the semantic annotation of places in location-based social networks. In *Proceedings of the 17th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 520–528). ACM.

Ye, M., Yin, P., Lee, W. C., & Lee, D. L. (2011). Exploiting geographical influence for collaborative point-of-interest recommendation. In *Proceedings of the 34th international ACM SIGIR Conference on research and development in information retrieval* (pp. 325–334). ACM.

Yin, H., & Cui, B. (2016). *Spatio-temporal recommendation in social media*. Springer.

Yin, H., Cui, B., Chen, L., Hu, Z., & Zhou, X. (2015). Dynamic user modeling in social media systems. *ACM Transactions on Information Systems (TOIS)*, *33*(3), 10.

Yin, H., Cui, B., Zhou, X., Wang, W., Huang, Z., & Sadiq, S. (2016). Joint modeling of user check-in behaviors for real-time point-of-interest recommendation. *ACM Transactions on Information Systems (TOIS)*, *35*(2), 11.

Yin, H., Sun, Y., Cui, B., Hu, Z., & Chen, L. (2013). Lcars: a location-content-aware recommender system. In *Proceedings of the 19th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 221–229). ACM.

Zhang, D., Chan, C. Y., & Tan, K. L. (2014). Processing spatial keyword query as a top-k aggregation query. In *Proceedings of the 37th international ACM SIGIR conference on research & development in information retrieval* (pp. 355–364). ACM.

Zhang, Y., Wang, H., Yin, G., Wang, T., & Yu, Y. (2017). Social media in GitHub: the role of@-mention in assisting software development. *Science China. Information Sciences*, *60*(3), 032102.

Zhao, K., Cong, G., Yuan, Q., & Zhu, K. Q. (2015). SAR: A sentiment-aspect-region model for user preference analysis in geo-tagged reviews. In *Data engineering, 2015 IEEE 31st international conference on* (pp. 675–686). IEEE.

Zhao, W. X., Jiang, J., Weng, J., He, J., Lim, E. P., Yan, H., et al. (2011). Comparing twitter and traditional media using topic models. In *European conference on information retrieval* (pp. 338–349). Springer.

Zhou, G., Yu, L., Zhang, C. X., Liu, C., Zhang, Z. K., & Zhang, J. (2015). A novel approach for generating personalized mention list on micro-blogging system. In *Data mining workshop, 2015 IEEE international conference on* (pp. 1368–1374). IEEE.