A Contextual Recurrent Collaborative Filtering Framework for Modelling Sequences of Venue Checkins

Jarana Manotumruksa[1], Craig Macdonald[2] and Iadh Ounis[2]
University of Glasgow
Glasgow, Scotland, UK
[1]j.manotumruksa.1@research.gla.ac.uk,[2]first.lastname@glasgow.ac.uk

Abstract

Context-Aware Venue Recommendation (CAVR) systems aim to effectively generate a ranked list of interesting venues users should visit based on their historical feedback (e.g. checkins) and context (e.g. the time of the day or the user's current location). Such systems are increasingly deployed by Location-based Social Networks (LBSNs) such as Foursquare and Yelp to enhance the satisfaction of the users. Matrix Factorisation (MF) is a popular Collaborative Filtering (CF) technique that can suggest relevant venues to users based on an assumption that similar users are likely to visit similar venues. In recent years, deep neural networks have been successfully applied to recommendation systems. Indeed, various approaches have been previously proposed in the literature to enhance the effectiveness of MF-based approaches by exploiting Recurrent Neural Networks (RNN) models to capture the sequential properties of observed checkins. Moreover, recently, several RNN architectures have been proposed to incorporate contextual information associated with the users' sequence of checkins (for instance, the time interval or the geographical distance between two successive checkins) to effectively capture such short-term preferences of users. In this work, we propose a Contextual Recurrent Collaborative Filtering Framework (CRCF) that leverages the users' preferred context and the contextual information associated with the users' sequence of checkins in order to model the users' short-term preferences for CAVR. In particular, the CRCF framework is built upon two state-of-the-art approaches: namely Deep Recurrent Collaborative Filtering framework (DRCF) and Contextual Attention Recurrent Architecture (CARA). Thorough experiments on three large checkin and rating datasets from commercial LBSNs demonstrate the effectiveness and

Authors version of article accepted in Information Processing and Management

robustness of our proposed CRCF framework by significantly outperforming various state-of-the-art matrix factorisation approaches. In particular, the CRCF framework significantly improves NDCG@10 by 5-20% over the state-of-the-art DRCF framework [1] and the CARA architecture [2] across the three datasets. Furthermore, the CRCF framework is less significantly risky than both the DRCF framework and the CARA architecture across the three datasets.

---

1. Introduction

Users in Location-Based Social Networks (LBSNs), such as Yelp and Foursquare, tend to search for interesting venues such as restaurants and museums to visit and can share their location with their friends by making checkins at the venues they have visited. This results in large amounts of user checkin data being received by the LBSNs. Such implicit feedback by users also provides rich information about both users and venues, and thus can be leveraged to study the users' movement in urban cities, as well as to enhance the quality of personalised venue recommendations. Effective Context-Aware Venue Recommendation systems (CAVRs) have become an essential application for LBSNs that allow users to find interesting venues based on their historical checkins and current context (e.g. time of the day, user's current location as well as their recently visited venues). Matrix Factorisation (MF) [3] is a Collaborative Filtering technique that is widely used to generate a personalised ranked list of venues to the users based on their historical checkins. In particular, the MF-based approaches for CAVR typically aim to embed the users' and venues' preferences as well as the contextual information about the users within *latent factors*, which are combined with a dot product operator to estimate the user's preference for a given venue and context.

Previous studies [4, 5, 6, 7, 8] have shown that the sequences of user's implicit feedback (e.g. sequences of checkins or clicks) play an important role in enhancing the effectiveness of recommendation, across various scenarios. However, traditional MF-based approaches can only capture users' long-term (*static*) preferences and not their short-term (*dynamic*) preferences. Here, *dynamic* preferences that are captured from the users' recently visited venues can influence the next venue they may visit (e.g. users may prefer to visit a bar directly after a dinner at a restaurant). In recent years, various approaches have been proposed to leverage Deep Neural Network (DNN) algorithms such

as Multi-Layer Perceptron (MLP), Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) for recommendation systems [9, 10, 4, 5, 11]. Among various DNN techniques, the RNN models have been widely exploited to extend the MF-based approaches to capture users' short-term preferences from their sequences of implicit feedback [12, 13, 5, 4, 14].

A common technique to incorporate RNN models (e.g. *Long Short-Term Memory* (LSTM) units [15] and *Gated Recurrent Units* (GRU) [16]) into MF-based approaches is to feed a sequence of user-venue interactions/checkins into a recurrent model and use the *hidden state* of the recurrent models to represent the users' dynamic preferences [4, 5, 6]. Next, the user's preference for a target venue is estimated by calculating the dot product between a latent factor of the user's dynamic preferences (i.e. the output of the recurrent models) and a latent factor [1] of the target venue. In addition, various approaches have been proposed to extend the RNN models to incorporate the contextual information associated with the sequences of user's implicit feedback for many recommendation tasks [12, 14, 13, 17, 18, 19, 20, 2]. Recently, Manotumruksa *et al.* [2] proposed a Contextual Attention Recurrent Architecture (CARA) that separately incorporates different types of contextual information associated with the users' sequence of implicit feedback to model the users' dynamic preferences for CAVR. The CARA architecture includes two gating mechanisms, namely a Contextual Attention Gate (CAG) and a Time- and Spatial-based Gate (TSG). The CAG controls the influence of context and the previous visited venues, while TSG controls the influence of the hidden state of the previous RNN unit, based on the time interval and the geographical distance between two successive checkins.

Similar to the RNN models proposed in the previous literature (e.g. [14, 13, 12], the CARA architecture still relies on a dot product of latent factors of users and items to capture the users' dynamic preferences in a Collaborative Filtering manner. However, previous works [11, 1] have shown that the dot product of latent factors may not be sufficient to capture the complex structures of the user-item interactions [10]. Recently, Manotumruksa *et al.* [1] proposed a Deep Recurrent Collaborative Filtering framework (DRCF) for venue recommendation that leverages the MLP and RNN models to learn the complex structures of the users' sequences of checkins by replacing the dot product with a neural architecture that can learn an ar-

[1] In this work, we use terms, latent factor and embedding vector, interchangeably.

bitrary function from the sequences of user' checkins. However, the DRCF framework still relies on the traditional RNN models that are not sufficiently flexible to incorporate the user's preferred context as well as the contextual information associated with the user's sequences of checkins.

Both the CARA architecture and the DRCF framework leverage the sequence of user' implicit feedback (i.e. sequences of checkins) to capture the users' *dynamic* preferences. A common challenge that arises when obtaining implicit feedback by observing checkins is that only positive feedback can be observed, and MF-based approaches trained on only positive feedback are likely to be biased to those positive instances. To address this challenge, various negative sampling approaches have been proposed [21, 22, 10, 23, 1]. For example, the BPR negative sampling approach proposed by Rendle *et al.* [21] uniformly and randomly selects venues that the users have not visited as negative instances. Recently, Manotumruksa *et al.* [1] proposed a sequence-based (*dynamic*) negative sampling approach that takes the sequential properties of checkins and the geographical location of venues into account to enhance the effectiveness of venue recommendation, as well as to alleviate the cold-start user problem. In this article, we aim to address a gap between two state-of-the-art factorisation- and RNN-based approaches (namely the DRCF framework and the CARA architecture) to capture the users' dynamic preferences when making context-aware venue recommendations, and thereby demonstrate that DRCF and CARA can be effectively combined for this task. Overall, our contributions are summarised below:

- We propose the Contextual Recurrent Collaborative Filtering framework (CRCF), an extension of the DRCF framework [1], which incorporates both the users' preferred context and the contextual information associated with the sequence of checkins to effectively capture the users' dynamic preferences for CAVR. Indeed, the original DRCF framework cannot incorporate the contextual information when generating venue recommendations. Moreover, we propose to integrate the state-of-the-art Contextual Attention Recurrent Architecture (CARA) [2] into our proposed CRCF framework to effectively capture the users' dynamic preferences.

- We propose to apply a novel sequence-based (*dynamic*) negative sampling approach proposed by Manotumruksa *et al.* [1] that takes the sequential properties of checkins as well as the geographical location of

venues into account to enhance the effectiveness of our CRCF framework. This is proposed in order to alleviate the cold-start user problem.

- We conduct thorough and comprehensive experiments on 3 large-scale real-world datasets, from Brightkite, Foursquare and Yelp, to demonstrate the effectiveness of our proposed CRCF framework for CAVR by comparing it with state-of-the-art venue recommendation approaches. Moreover, we investigate the robustness of the CRCF framework by leveraging risk analyses techniques proposed by Wang *et al.* [24] and Dinçer *et al.* [25].

The experimental results presented in Section 6 demonstrate that our proposed CRCF framework consistently and significantly outperforms various state-of-the-art venue recommendation approaches in terms of effectiveness and robustness. In particular, our experimental findings are as follows:

- The contextual information associated with the sequences of the users' checkins (e.g. the time interval and distance between two successive checkins) is important in enhancing the quality of context-aware venue recommendation. Our proposed CRCF framework, which leverages the contextual information can significantly outperform both the state-of-the-art DRCF framework and the CARA architecture on three large datasets.

- Leveraging the sequential order of users' checkins as well as the geographical information of venues can enhance both the effectiveness and robustness of the CRCF framework. In particular, our experimental results show that the *dynamic* geo-based negative sampling approach, which takes into account both the sequential order of users checkins and the geographical information of venues, can significantly improve the effectiveness and robustness of various approaches (i.e. the DRCF and CRCF frameworks and the CARA architecture) as well as alleviate the cold-start problem.

- Throughout our comprehensive robustness analysis experiments, we observe that our proposed CRCF framework is significantly less risky, and is less likely to generate poor venue suggestions to the users across the three used datasets, compared to the DRCF framework and the CARA architecture. Moreover, the CRCF framework is more robust

than various state-of-the-art venue recommendation approaches (i.e. less likely to generate worse venue suggestions compared to a traditional CF baseline such as BPR).

This article is structured as follows: Section 2 provides the background literature on CAVR, recent trends in applying Deep Neural Networks to recommendation systems as well as various existing extensions of RNN models; Section 3 provides a brief description of the DRCF framework. Section 4 details how to extend the DRCF framework for the CAVR task and also how to integrate the CARA architecture into the resulting CRCF framework; Experimental setup and results are provided in Sections 5 & 6, respectively. Concluding remarks follow in Section 7.

2. Background

Context-Aware Venue Recommendation (CAVR). Collaborative Filtering (CF) techniques such as Matrix Factorisation (MF) [3], Factorisation Machines [7] and Bayesian Personalised Ranking (BPR) [21] have been widely used in recommendation systems. Such factorisation-based approaches assume that users who have visited similar venues share similar preferences, and hence are likely to visit similar venues in the future. Previous works on venue recommendation have shown that the contextual information associated with the users' observed feedback (time of the day, location) plays an important role to enhance the effectiveness of CAVR as well as to alleviate the cold-start problem [23, 26, 22, 27, 28, 29, 30]. For example, Yao *et al.* [26] extended the traditional MF-based approach by exploiting a high-order tensor instead of a traditional user-venue matrix to model multi-dimensional contextual information. Manotumruksa *et al.* [23] and Yuan *et al.* [22] extended BPR to incorporate the geographical location of venues to alleviate the cold-start problem by sampling negative venues based on an assumption that users prefer nearby venues over distant ones. Zhao *et al.* [30] proposed Spatial-TEmporaL LAtent Ranking (STELLAR), which recommends a list of venues based on the user's context such as time and recent checkins. However, similar to traditional MF, these approaches still rely on the dot product operation (i.e. a linear function) to estimate the users' preferences from their latent factors, which has been recently demonstrated to be less effective than non-linear activation functions [1, 10]

Deep Neural Network Recommendation Systems. With the impressive successes of Deep Neural Network (DNN) models in domains such as speech recognition, computer vision and natural language processing (e.g. [31, 32, 33]), various approaches (e.g. [4, 1, 10, 11, 9, 19, 34]) have been proposed to exploit DNN models for recommendation systems. For example, He *et al.* [10] and Cheng *et al.* [9] proposed to exploit Multi Layer Perceptron (MLP) models to capture the complex structure of user-item interactions. An advantage of such MLP-based models is their ability to capture the user's complex structure using a DNN architecture and a non-linear activation function such as the sigmoid function. Liu *et al.* [34, 19] and Manotumruksa *et al.* [1] all exploited Recurrent Neural Networks (RNNs) to model the sequential order of the users' observed feedback. In particular, Manotumruksa *et al.* [1] proposed a Deep Recurrent Collaborative Filtering (DRCF) framework that captures the users' short-term preferences from their sequence of checkins by exploiting an RNN model. We use DRCF as the basis of our proposed framework in this work. Due to the complex and overwhelming number of parameters of DNN models (i.e. the MLP and RNN models), these existing DNN-based CF approaches are likely to be prone to overfitting. Several empirical studies [11, 10, 35] have demonstrated that the use of generalised distillation techniques, such as dropout & regularisation, as well as pooling techniques can alleviate the overfitting problems inherent in DNN-based CF approaches.

The previous work mentioned above mainly focused on how to exploit existing DNN models to enhance the quality of recommendations. However, fewer attempts have addressed how to extend such DNN models to address particular challenges in recommendation systems. In particular, few approaches have been proposed to extend the traditional RNN models (e.g. Long Short-Term Memory (LSTM) [15] and Gated Recurrent Units (GRU) [16]) to incorporate the contextual information of observed feedback into various recommendation settings (e.g. [12, 14, 13, 18, 2]). In this vein, we note Zhu *et al.* [12], who proposed an extension of LSTM (TimeLSTM) by introducing time gates that control the influence of the hidden state of a previous LSTM unit based on the time interval between successive observed feedback. Indeed, they assumed that the shorter the time interval between two successive feedback, the stronger the correlation between these two feedback. Smirnova and Vasile [14] proposed a Contextual RNN architecture that can incorporate different types of context that are observed from the users' checkins (e.g. user's current location and the time of the day). Build-

ing upon Smirnova and Vasile's work [14], Beutel *et al.* [13] explored various approaches to effectively incorporate the latent factors of context into RNN models. They proposed LatentCross, a technique that incorporates contextual information within a GRU, by performing an element-wise product of the latent factors of context with the model's hidden states. Recently, based on TimeLSTM [12] and LatentCross [13], Manotumruksa *et al.* [2] proposed a Contextual Attention Recurrent Architecture (CARA), an extension of the GRU architecture that incorporates the contextual information associated with the sequence of users' feedback to generate effective context-aware venue recommendations.

In the next section, we provide an overview of the Deep Recurrent Collaborative Filtering (DRCF) framework. In particular, we describe how the DRCF framework leverages the sequential order of the users' checkins to capture their short-term (*dynamic*) preferences. Later, in Section 4, we explain how to extend the DRCF framework to incorporate the contextual information associated with the users' sequence of checkins to generate effective context-aware venue recommendations, by integrating the CARA RNN architecture.

3. Deep Recurrent Collaborative Filtering framework (DRCF)

In this section, we first formalise the problem statement as well as the notations used in this article (Section 3.1). Then, we briefly describe the DRCF framework for venue recommendation in Section 3.2. Later in Section 4, we describe in detail our proposed Contextual Recurrent Collaborative Filtering (CRCF) framework, an extension of the DRCF framework that incorporates the contextual information associated with the sequences of user's checkins to enhance the quality of CAVR.

3.1. Problem Statement

The task of context-aware venue recommendation is to generate a ranked list of venues that a user might visit given his/her preferred context and historical feedback (e.g. the previously visited venues from checkin data). Let $c_{i,j,t}$ denote a user $i \in \mathcal{U}$ who has made a checkin to venue $j \in \mathcal{V}$ at timestamp t. Note that $c_{i,j,t} = 0$ means that user i has not made a checkin at venue j at time t. Let $\mathcal{V}_i^+$ denote the list of venues that the user i has previously visited, sorted by time, and let $\mathcal{S}_i$ to denote the set of sequences of checkins

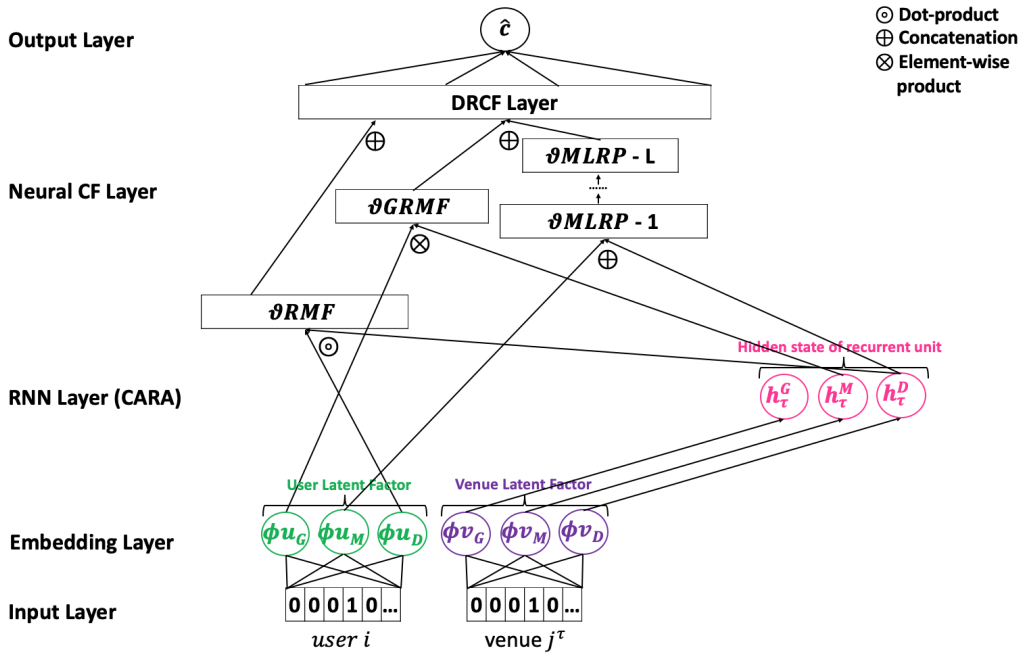


Figure 1: A diagram of the Deep Recurrent Collaborative Filtering (DRCF) framework.

(e.g. $\mathcal{S}_i = \{[c_1], [c_1, c_2], [c_1, c_2, c_3]\}$). $s_{i,t} = \left\{c = (i, j, \grave{t}) \in \mathcal{C} \mid \grave{t} < t\right\} \subset \mathcal{S}_i$ denotes the sequence of checkins of user i up to time t. We use $s^{\tau}_{i,t}$ to denote the τ-th checkin in the sequence $s_{i,t}$. t^{τ} denotes the timestamp of τ-th checkin. Finally, lat_j, lng_j are the latitude and longitude of checkin/venue j.

3.2. Deep Recurrent Collaborative Filtering framework (DRCF)

The DRCF framework proposed by Manotumruksa *et al.* [1] is illustrated in Figure 1. It consists of five layers with the connections between the layers represented using red-dashed lines. Starting at the bottom of the figure, at time step τ, the input layer consists of a binary sparse vector with a one-hot encoding that represents user i and venue j, respectively. The sparse vectors of the user and venue are fed into the embedding layer. In the embedding layer, there are the embedding representations for users and venues, highlighted in green and purple, respectively. The outputs of the embedding layer can be seen as the latent factors of each user and venue (respectively we denote these as $\phi u_i \in \mathbb{R}^d$, $\phi v^{\tau}_j \in \mathbb{R}^d$, where d is the number of latent dimensions). $\theta_e = \{\phi u_D \in \mathbb{R}^{|\mathcal{U}| \times d}, \phi u_M \in \mathbb{R}^{|\mathcal{U}| \times d}, ..., \phi v_D \in \mathbb{R}^{|\mathcal{V}| \times d}\}$ denotes the set of parameters of the embedding layers. Next, the latent factors of the venues

are fed into the Recurrent Neural Networks (RNN). In the RNN layer, DRCF exploits the traditional RNN models to encapsulate the users' *dynamic* preferences from their sequence of checkins. The outputs of the RNN layer is a hidden state of recurrent unit $h_\tau = \sigma(X\phi v_j^\tau + Rh_{\tau-1})$ that represents the *dynamic* preferences of user i at time step τ. $\sigma(x) = 1/(1+e^{-x})$ is the sigmoid function. R is a recurrent connection weight matrix that captures the sequential signals between every two adjacent hidden states $h_{\tau-1}$ and h_τ while X is a transition matrix between the latent factors of venues. By $\theta_r = \{R, X\}$, we denote the set of parameters of the RNN model. Next, the user's *dynamic* preferences h_τ, and the latent factors of user i, ϕi, are fed into the Neural Collaborative Filtering layers, which consist of the Generalised Recurrent Matrix Factorisation (GRMF), the Multi-Level Recurrent Perceptron (MLRP) and Recurrent Matrix Factorisation (RMF) models, to discover certain latent structures of sequences of user-venue interactions. The outputs of these models are concatenated and fed into the output layer. Finally, the output layer provides a score predicting if the user i will checkin at venue j, which is defined as follows:

$$\hat{c}_{i,j} = \sigma(H(\vartheta^{GRMF} \oplus \vartheta^{MLRP} \oplus \vartheta^{RMF})) \tag{1}$$

where $\oplus$ denotes the concatenation operation. line:s:notation$\vartheta^{GRMF} = h_\tau^G \otimes \phi u_{iG}$ is the output of the GRMF model, where $\otimes$ denotes the element-wise product operation. $\vartheta^{MLRP} = a_L(H_L(...a_1(H_1(h_\tau^M \oplus \phi u_{iM}))))$ is the output of the MLRP model, where L is a number of layers and a denotes the activation function. Following Manotumruksa *et al.* [2], we use a rectified linear unit ($ReLU(x) = max(0, x)$) as the activation function a. $\vartheta^{RMF} = h_\tau^D \circledcirc \phi u_{i_D}$ is the output of the RMF model, where $\circledcirc$ denotes the dot product operation.line:e:notation $H(x) = (W^T x + b)$ is the hidden layer, where W and b are the weight matrix and bias vector, respectively. Overall, $\theta_H = \{W, b\}$ denotes the set of parameters of the hidden layers. $H(x)$ ensures that each dimension of the latent factors from ϑ^{GRMF}, ϑ^{MLRP} and ϑ^{RMF} are independent (i.e. each dimension of the latent factors are treated independently by the hidden layers). Note that the DRCF framework exploits different embedding and RNN layers for each model in order to independently learn the complex structures of user-venue interactions from different models (i.e. the GRMF, MLRP and RMF models capture the interactions using the element-wise product, concatenation and dot product operations, respectively). The benefit of the DRCF framework is that it allows different models to be learned from different sets of embeddings and RNN layers, hence capturing different

characteristics of the task. Indeed, while these embeddings and RNN layers have been explored in the literature [10, 1], later in Section 4, we extend the DRCF framework to incorporate the contextual information associated with the users' sequence of checkins by exploiting the state-of-the-art RNN architecture.

Next, the DRCF framework applies the Bayesian Personalised Ranking (BPR) model to learn the parameters $\Theta = \{\theta_r, \theta_e, \theta_h\}$. Note that the BPR model consists of a pairwise ranking function and a negative sampling process. As mentioned before, with the implicit feedback in the form of checkins, only the users' activities are observed (i.e. users have visited some venues at some particular times), while their preferences on those venues cannot be observed. To alleviate this problem, inspired by the negative sampling process of the BPR model, for each user, given the sequence of the user's checkins, DRCF samples, as negative instances, venues that the user has never visited before, $\mathcal{V} - s_{i,t}$. Then, DRCF aims to rank the venues that the users have previously visited higher than the unvisited venues. In particular, the objective function of the DRCF framework is defined as follows:

$$\mathcal{J}(\Theta) = \sum_{i \in \mathcal{U}} \sum_{s_{i,t} \in S_i} \sum_{(i,j,t) \in s_{i,t}} \sum_{k \in \mathcal{V} - s_{i,t}} \log(\sigma(\hat{c}_{i,j} - \hat{c}_{i,k})) \tag{2}$$

where j is the venue most recently visited in s_t, k is an unvisited venue. Next, Manotumruksa *et al.* [1] proposed a *dynamic* geo-based negative sampling approach to enhance the effectiveness of the DRCF framework as well as alleviate the cold-start problem. In particular, they modified the objective of the DRCF framework (Equation (2)) to incorporate the geographical information of venues during the sampling process as follows:

$$\mathcal{J}(\Theta) = \sum_{i \in \mathcal{U}} \sum_{s_{i,t} \in S_i} \sum_{(i,j,t) \in s_{i,t}} \sum_{k \in \mathcal{N}_j - s_{i,t}} \sum_{l \in \mathcal{V} - s_{i,t}} \Bigg[\log(\sigma(\hat{c}_{i,j} - \hat{c}_{i,k})) - \log(\sigma(\hat{c}_{u,k} - \hat{c}_{u,l})) \Bigg] \tag{3}$$

line:s:samplingwhere j is the venue most recently visited in s_t, k is an unvisited venue that is nearby to venue j, l is an unvisited venue that is far away from venue j and $\mathcal{N}_j$ is the set of venues that are nearby to venue j. Intuitively, this *dynamic* geo-based negative sampling approach assumes that the nearby unvisited venue k should be ranked higher than distant and

unvisited venue l because the users are likely to visit new venues nearby to the venues they previously visited. By leveraging the geographical locations of venues during the sampling process, we can alleviate the cold-start problem by effectively sampling the nearby yet unvisited venues as negative instances.line:e:sampling

In this section, we have provided an overview of the Deep Recurrent Collaborative Filtering (DRCF) framework. In particular, we described how the DRCF framework leverages the sequential *order* of the users' checkins to capture their short-term (*dynamic*) preferences. In the next section, we explain how to extend the DRCF framework to incorporate the *contextual* information associated with the users' sequence of checkins (such as the time interval and the distance between successive checkins) to generate effective context-aware venue recommendations

4. Contextual Recurrent Collaborative Filtering Framework (CRCF)

In this section, we describe a Contextual Recurrent Collaborative Filtering framework (CRCF), an extension of the DRCF framework, that can effectively incorporate different types of contextual information associated with the sequential feedback (i.e. the time interval and geographical distance between two successive checkins) to model users' short-term (dynamic) preferences. In particular, the CRCF framework aims to generate a ranked-list of venues that a user might prefer to visit at time t based on the sequences of checkins $s_{i,t}$. The CRCF framework consists of five layers - the connections between these layers are presented using both blue- and red-dashed lines in Figure 2. The structure of the CRCF framework is different from the structure of the DRCF framework including its input, embedding and RNN layers. Starting at the bottom of the figure, at the input layer, at time step τ, given a user i, venue j and time t^τ, we compute the time interval and the geographical distance between the given venue j and the venue k, which was previously visited at time step $\tau - 1$, as $\Delta t^\tau = t^\tau - t^{\tau-1}$ and $\Delta g_\tau = dist(lat_j, lng_j, lat_k, lng_k)$, respectively. $dist()$ is the Haversine distance function that returns the distance between the given latitudes and longitudes. In the embedding layer, there are three adding layerighted in yellow in Figure 2 that are used to generate the latent factors of the time $\phi t^\tau \in \mathbb{R}^d$. Note that we only consider the time of checkins as the user's preferred context. However, our proposed framework is sufficiently flexible to support other possible types of context (e.g. the current weather of the day).

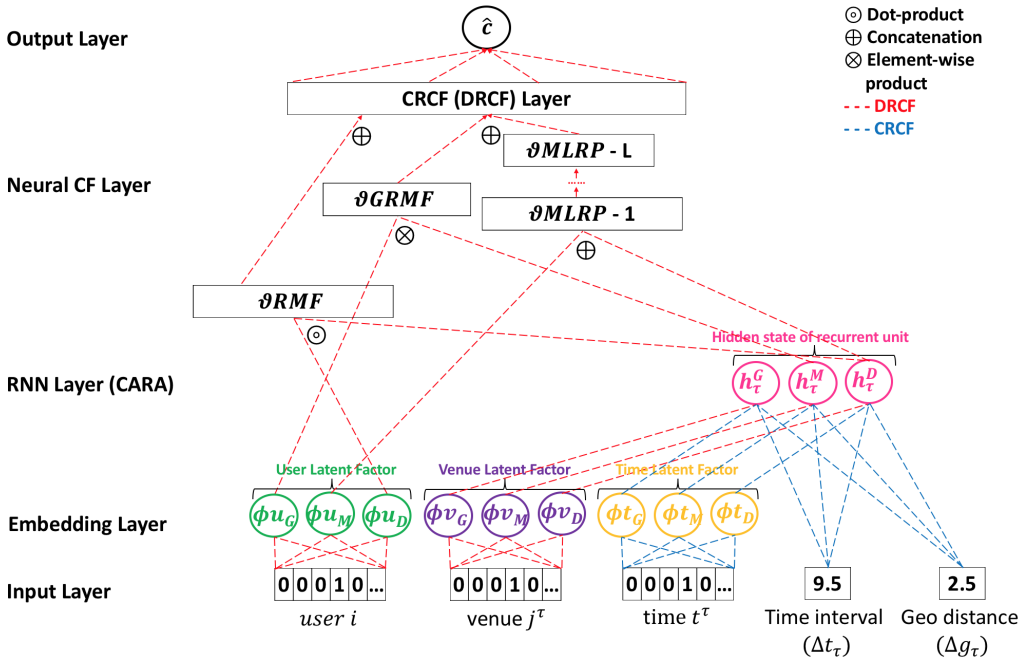


Figure 2: A diagram of the Contextual Recurrent Collaborative Filtering (CRCF) framework. The connections of each layer linked by the red-dashed lines illustrate the DRCF framework. The connections of each layers linked by the blue-dashed and red-dashed lines illustrate the CRCF framework, which is an extension of the DRCF framework.

Next, the latent factors of venue and time (ϕv_j^τ and ϕt^τ) as well as the time interval Δt_τ and the geographical distance Δg_τ are fed into the RNN layer. In the RNN layer, we exploit the CARA architecture proposed by Manotumruksa *et al.* [2] rather than the traditional RNN models used by the DRCF framework to encapsulate the *dynamic* user preferences. In particular, the main advantage of the CARA architecture over the traditional RNN models is that it can effectively capture the users' *dynamic* preferences by taking the contextual information associated with the users' two successive checkins into account. The output of the recurrent layer is the hidden state of the recurrent unit at time step τ, $h_\tau \in \mathcal{R}^d$, which is defined as follows:

$$h_\tau = f_{CARA}(\phi v_j^\tau, \phi t^\tau, \Delta t_\tau, \Delta g_\tau; \theta_r) \tag{4}$$

where $\theta_r = \{W, R, U, b\}$ denotes the set of parameters of the recurrent layer. Further details of the CARA architecture, f_{CARA}, are described in [2]. Then, similar to the DRCF framework, the latent factors of user ϕu_i, and the user's

dynamic preferences h_τ are fed into the Neural CF layer and the output layer, respectively. The objective function of the CRCF framework is similar to DRCF's, as described in Equation (3).

There are two advantages of the CRCF framework over either the DRCF framework or the CARA architecture. First, CRCF allows to take the user's context into account to generate effective venue recommendations based on his/her context, while DRCF cannot. Although CARA can incorporate the user' context during the recommendation process, it still relies on the dot product of the latent factors when making recommendations. Indeed, previous works [10, 1] have shown that the dot product operation is not effective in capturing the complex structure of user-venue interactions. Unlike CARA, our proposed CRCF approach is built upon the DRCF framework, which exploits the element-wise product and the concatenation operation to effectively capture the complex structure of user-venue interactions.

5. Experimental Setup

In this section, we evaluate the effectiveness and robustness of our proposed Contextual Recurrent Collaborative Filtering (CRCF) framework in comparison with various matrix factorisation-based approaches. In particular, we aim to address the following research questions:

RQ1 *Can we enhance (a) the effectiveness and (b) the robustness of the Contextual Recurrent Collaborative Filtering (CRCF) framework for CAVR, by exploiting the state-of-the-art Contextual Attention Recurrent Architecture (CARA) to leverage the time interval and the geographical distance associated with sequences of checkins?*

RQ2 *Can the* dynamic *geo-based negative sampling approach proposed by Manotumruksa* et al. *[1], which leverages both the sequential properties of checkins and the geographical location of venues, enhance (a) the effectiveness and (b) the robustness of CRCF and alleviate the cold-start problems?*

In the remainder of this section, we describe the experimental setup in terms of datasets (Section 5.1), baselines (Section 5.2), algorithm parameters (Section 5.3) and measures(Section 5.4). The experimental results and analysis follow in Section 6.

Table 1: Statistics of the three used datasets.

	Brightkite	Foursquare	Yelp
Number of normal users	14,374	10,766	38,945
Number of venues	5,050	10,695	34,245
Number of ratings or checkins	681,024	1,336,278	981,379
Number of cold-start users	5,578	154	6903
% density of User-Venue matrix	0.93	1.16	0.07

5.1. Datasets

We conduct experiments using three publicly available large-scale user-venue interaction datasets from LBSNs. In particular, to show the generalisation of our proposed CRCF framework across multiple LBSN platforms and sources of feedback evidence, we use two checkin datasets from Brightkite[2] and Foursquare[3], and a rating dataset from Yelp[4]. Following the common practice from previous works [21, 1, 10, 2], we remove venues with less than 10 checkins. Table 1 summarises the statistics of the filtered datasets. To evaluate the effectiveness of our proposed CRCF framework and following previous works [10, 1, 2], we adopt a *leave-one-out* evaluation methodology: for each user, we select his/her most recent checkin as a ground truth and randomly select another 100 venues that the user has not visited before as the testing set, where the remaining checkins are used as the training and validation sets. The context-aware venue recommendation task is thus to rank those 101 venues for each user, given their preferred context (i.e. time), aiming to rank the highest the most recent ground truth checkin. line:s:taskNote that the context-aware venue recommendation task allows to recommend venues that the user has previously visited, for example in a different context. For instance, while a user may have visited a restaurant a week ago, recommending the same restaurant to the user to visit in the next few hours is acceptable.line:e:task

We conduct two separate sets of experiments, namely: *Normal Users* (those with ≥ 10 checkins) and *Cold-start Users* (< 10 checkins) to evaluate the effectiveness of our proposed CRCF framework in the general and cold-start settings.

5.2. Baselines

We compare our proposed Contextual Recurrent Collaborative Filtering (CRCF) framework with various matrix factorisation-based approaches. We

[2] `https://snap.stanford.edu/data/` [3] `https://archive.org/details/201309_foursquare_dataset_umn`
[4] `https://www.yelp.com/dataset_challenge`

Table 2: Summary of baselines.

	MF [3]	BPR [21]	GeoBPR [22]	STELLAR [30]	NeuMF [10]	DRCF [1]	RNN [6]	DREAM [4]	CARA [2]
Neural networks	×	×	×	×	✓	✓	✓	✓	✓
Sequential-based	×	×	×	✓	×	✓	✓	✓	✓
Context-aware	×	×	only geo	✓	×	×	×	×	✓
Ordinary/Transition	×	×	×	✓	×	×	×	×	✓
Special gates	×	×	×	×	×	×	×	×	✓

implement all baselines and the CRCF framework using Keras[5], a deep learning framework built on top of Theano[6]. Our implementation of the CRCF framework is available as open source[7]. Note that some baselines may not have been originally proposed for venue recommendation but are sufficiently flexible to be applied to such a task without any disadvantage. The choice of recurrent models is fixed to the GRU units proposed by Zhang *et al.* [16]. Table 2 characterises the various baselines into different categories, namely neural network-based approaches, sequential-based approaches, context-aware based approaches, approaches that take both the *ordinary* and *transition* context into account, as well as approaches that make use of adapted RNN gates. All baselines are summarised below: line:s:baseline

MostPop. A baseline that ranks venues in descending order of the venues' popularities, calculated across all users.

MostVisit. This baseline ranks venues for a given user in descending order of the venues' popularity for that user.

RecentVisit. A baseline that takes the user's sequential order of checkins into account and recommends the most recently visited venue to the user. line:e:baseline

MF. The traditional matrix factorisation approach proposed by Koren *et al.* [3] that aims to accurately predict the users' checkin on the unvisited venues.

BPR. The classical pairwise ranking approach, coupled with matrix factorisation for user-venue checkin prediction, proposed by Rendle *et al.* [21].

GeoBPR. An extension of BPR that incorporates geographical location of venues to sample negative venues that are far away from the user's previous visits. GeoBPR was proposed by Yuan *et al.* [22].

[5] `https://github.com/fchollet/keras` [6] `http://deeplearning.net/software/theano`
[7] `https://github.com/feay1234/CRCF`

RNN. A sequential click prediction with recurrent neural networks approach proposed by Zhang *et al.* [6].

DREAM [4]. A RNN model that incorporates BPR for ranking optimisation. As DREAM is originally proposed for next shopping-basket recommendation, to permit a fair comparison with our proposed DRNN approach, we reimplement DREAM to treat a single checkin as the shopping-basket purchase.

NeuMF. A Neural Matrix Factorisation framework[8], proposed by He *et al.* [10], which exploits Multi-Level Perceptron (MLP) models to capture the complex structure of user-item interactions.

DRCF. A Deep Recurrent Collaborative Filtering framework for venue recommendation proposed by Manotumruksa *et al.* [1], which extends NeuMF [10] to exploit the RNN-based models to model the sequences of users' checkins (see Section 3).

STELLAR. A Spatial-TEmporaL LAtent Ranking framework for CAVR proposed by Zhao *et al.* [30] that aims to recommend the list of venues based on the user's preferred time and last successive visits. Note that this is the only context-aware framework that does not rely on the RNN-based approaches to model the users' sequential order of checkins.

CARA. A state-of-the-art Contextual Attention Recurrent Architecture[9] for CAVR proposed by Manotumruksa *et al.* [2] that leverages the contextual information associated with the sequence of user's checkins to model the user's dynamic preferences.

5.3. Recommendation Parameter Setup

Following [1, 10, 2], we set the dimension of the latent factors d and hidden layers h_τ of our proposed CARA architecture and all of the matrix factorisation-based approaches to be identical: $d = 10$ across three datasets. Later, in Section 6.1.1, we vary the dimension of the latent factor to empirically verify their impact on effectiveness. Following He *et al.* [10], we randomly initialise all embeddings and recurrent layers' parameters, $\theta_r, \theta_e, \theta_h$, with a Gaussian distribution (with a mean of 0 and a standard deviation of 0.01) and apply the mini-batch Adam optimiser [36] to optimise those

[8] `https://github.com/hexiangnan/neural_collaborative_filtering`
[9] `https://github.com/feay1234/CARA`

parameters, which yields a faster convergence than SGD and automatically adjusts the learning rate for each iteration. We initially set the learning rate to 0.001[10] and set the batch size to 256. Since the impact of the recurrent parameters such as the size of the hidden state, have been explored in previous works [10, 37, 35], in this article we omit varying the size of the hidden layers and the dimension of the latent factors. Indeed, larger sizes of hidden layers and dimensions may cause overfitting and degrade the generalisation of the models [37, 10, 35].

5.4. Measures

We measure the quality of the ranked list of venues in terms of Hit Ratio (HR) and Normalized Discounted Cumulative Gain (NDCG) – as applied in the existing literature [10, 4, 1, 2]. In particular, HR considers the ranking nature of the task, by taking into account the rank(s) of the venues that each user has previously visited/rated in the produced ranking, while NDCG goes further by considering the checkin frequency/rating value of the user as the graded relevance label. Finally, significance tests are conducted using a paired t-test.

Furthermore, we experiment to determine the *robustness* of the CRCF framework, to measure its likelihood to underperform in comparison to an established baseline recommender system. Throughout our robustness experiments, we use the Bayesian Personalised Ranking (BPR) model, which is equivalent to BM25 baseline in web search, as the established baseline for venue recommendation system to evaluate the robustness of our proposed CRCF framework. To this end, we use risk-sensitive evaluation measures to quantify any underperformance compared to a given baseline model (i.e. the BPR model). All risk-sensitive measures are defined in terms of Risk & Reward [24], where Risk is defined as the average reduction in effectiveness due to the use of the new target model in comparison to the baseline CF ranking model. In contrast, Reward is the positive improvement in effectiveness of the target model over the baseline model, averaged across all users. We use NDCG as the primary effectiveness measure for comparing the effectiveness of the new target model and the baseline CF ranking model. In particular, given a baseline CF ranking model (i.e. BPR), the Risk and Reward scores of using a target model (e.g. DRCF or CRCF) over the set of all users are

[10] The default learning rate setting of the Adam optimiser in Keras.

measured as follows:

$$Reward = \frac{1}{|\mathcal{U}|} \sum_{i \in \mathcal{U}} max(0, M_t(i) - M_b(i)) \tag{5}$$

$$Risk = \frac{1}{|\mathcal{U}|} \sum_{i \in \mathcal{U}} max(0, M_b(i) - M_t(i)) \tag{6}$$

where M_b and M_t denote the effectiveness of the baseline CF ranking model and the target model for a given user i, respectively, calculated using NDCG. Let the overall gain of a target model be $Gain = reward - risk$. Next, a single measure that takes the risk-reward tradeoff into account is calculated as $U_{risk} = Gain - \alpha \cdot Risk$, where $\alpha \geq 0$ is a risk-sensitivity parameter [24]. Note that with $\alpha = 0$, U_{risk} simply measures the average difference in performances between the two models across all users; On the other hand, increasing $\alpha > 0$ places more emphasis on penalising models that underperform compared to the baseline.

Following Dinçer *et al.* [25], for $\alpha \geq 0$, a t-statistic can be formulated based on U_{risk}, which they call T_{risk}, and can be expressed as follows:

$$T_{risk} = \frac{U_{risk}}{SE(U_{risk})} \tag{7}$$

where $SE()$ is the standard error of the paired sample mean. The advantage of T_{risk} over U_{risk} is that it is easily interpreted for an inferential analysis of risk (i.e. if the system exhibits a significant level of risk for a given α). Indeed, $\mathrm{T}_{risk} < 2$ denotes a significant risk [25] at $p < 0.05$. Later in Section 6.2, we test the significance of an observed risk-reward tradeoff score between a target model and a given baseline by using T_{risk} as the test statistic of the Student's t-test for matched pairs[11].

6. Experimental Results

In this section, we report the effectiveness and robustness of our proposed CRCF framework in comparison with various state-of-the-art approaches. In particular, to address research questions RQ1(a) and RQ2(a), we conduct

[11] Note that for $\alpha = 0$, T_{risk} exceeding ± 2 simply denote a significant difference according the normal t-test with $p < 0.05$.

various experiments to evaluate the effectiveness of the CRCF framework under the *Normal* and *Cold-Start* settings, which are discussed in Section 6.1. Moreover, to answer research questions RQ1(b) and RQ2(b), we further perform several risk analysis experiments to investigate the robustness of the CRCF framework, which are discussed in Section 6.2

6.1. Effectiveness Evaluation

In this section, we report the effectiveness of our proposed CRCF framework in comparison with various state-of-the-art approaches. This section is structured to separately address research questions RQ1(a) and RQ2(a). In particular, to answer research question RQ1(a), Section 6.1.1 reports the performance of the CRCF framework and the used baselines under the the *Normal* and *Cold-Start* settings. Where negative sampling is used (e.g. DRCF and CRCF), we apply the traditional BPR negative sampling approach (Equation (2)). To answer research question RQ2(a), Section 6.1.2 demonstrates the usefulness of the *dynamic* geo-based negative sampling approach (Equation (3)) in enhancing the effectiveness of the CRCF framework and alleviating the cold-start problem.

line:s:sec1

6.1.1. Effectiveness of the CRCF framework

Table 3 reports the effectiveness of the CRCF framework in comparison with various matrix factorisation-based approaches in term of the HR@10 and NDCG@10 measures on the three used datasets. In particular, the table contains two groups of rows, which report the effectiveness of various approaches under the *Normal Users* and *Cold-Start Users* experiments, respectively. Similar to Table 3, Table 4 reports the observed performances of the CRCF and DRCF frameworks as well as the CARA architecture when incorporating the *dynamic* geo-based negative sampling approach (Equation (3)), which takes the geographical location of venues into account during the negative sampling process.

Firstly, on inspection of the first group of rows of Table 3, we note that the relative venue recommendation quality of the baselines on the three datasets in terms of the two measures are consistent with the results reported for the various baselines in the corresponding literature [10, 4, 6, 1, 2]. For instance, DRCF outperforms MF, BPR and NeuMF across the three datasets. Similarly, CARA outperforms STELLAR across the three datasets. Note that previous works (i.e. NeuMF [10], DREAM [4],STELLAR [30]) used

Table 3: Performance in terms of HR@10 and NDCG@10 between various approaches. The best performing approach is highlighted in bold; − and ∗ denote a significant difference compared to the best performing result, according to the paired t-test for $p < 0.05$ and $p < 0.01$, respectively.

Normal Users Experiments						
	Brightkite		Foursquare		Yelp	
Model	HR	NDCG	HR	NDCG	HR	NDCG
MostPop	0.1462*	0.1010*	0.2009*	0.1167*	0.0739*	0.0334*
MostVisit	0.4032*	0.3473*	0.4733*	0.4290*	0.1083*	0.0528*
RecentVisit	0.4809*	0.4370*	0.4584*	0.4037*	0.1096*	0.0542*
MF	0.6206*	0.3470*	0.6656*	0.3818*	0.3539*	0.1734*
RNN	0.6368*	0.3824*	0.8040*	0.5459*	0.3814*	0.1891*
BPR	0.6890*	0.4333*	0.7550*	0.4834*	0.4963*	0.2676*
DREAM	0.7041*	0.4839*	0.8147*	0.6081*	0.4349*	0.2235*
STELLAR	0.7267*	0.5635*	0.8751*	0.6984*	0.5356*	0.2969*
NeuMF	0.7073*	0.5358*	0.8361*	0.5842*	0.4934*	0.2729*
DRCF	0.7419*	0.6048*	0.8952*	0.7223*	0.5162*	0.2963*
CARA	0.7385*	0.6040*	0.8851*	0.7154*	0.5587*	0.3272*
CRCF	**0.7528**	**0.6319**	**0.8981**	**0.7442**	**0.5861**	**0.3479**
Cold-Start Users Experiments						
	Brightkite		Foursquare		Yelp	
Model	HR	NDCG	HR	NCDG	HR	NDCG
MostPop	0.1155*	0.0778*	0.0584*	0.0286*	0.0714*	0.0316*
MostVisit	0.4285*	0.3789*	0.3506*	0.3175*	0.1044*	0.0489*
RecentVisit	0.4995*	0.4585*	0.3831*	0.3446*	0.1052*	0.0497*
MF	0.6768*	0.3913*	0.6623*	0.3650*	0.3748*	0.1868*
BPR	0.7519*	0.4907*	0.7792-	0.4961*	0.5273*	0.2946*
RNN	0.6486*	0.3694*	0.5909*	0.4041*	0.3856*	0.1901*
DREAM	0.7452*	0.4969*	0.7987-	0.5379*	0.4523*	0.2239*
STELLAR	0.7406*	0.5580*	0.8052-	0.6007*	0.5537*	0.3147*
NeuMF	0.7160*	0.5894*	0.7922-	0.6227*	0.5102*	0.2734*
DRCF	0.7526*	0.5980*	0.8377	0.6645	0.5330*	0.3136*
CARA	0.7648*	0.6220*	**0.8636**	0.6505-	0.5748*	0.3493*
CRCF	**0.7782**	**0.6582**	0.8571	**0.6967**	**0.5913**	**0.3622**

different datasets, but our reimplementations of their proposed approaches obtain similar relative improvements.

Comparing CRCF with the various baselines, we observe that CRCF consistently and significantly outperforms all baselines for both HR and NDCG, across all datasets. In particular, comparing with DRCF and CARA, CRCF obtains 4.61%, 3-4.02% and 6.32-17.41% improvements in terms of NDCG for Brightkite, Foursquare and Yelp datasets, respectively. These results suggest that our proposed framework, an extension of DRCF that exploits the CARA architecture instead of the traditional RNN models to leverage the user's preferred context (i.e. time) and the contextual information associated with the sequence of checkins, is more effective than the DRCF framework, which ignores those contexts. Moreover, comparing CRCF with the CARA architecture, which both take the users' context into account, the results suggest

that the neural architecture in the CRCF framework (i.e. an element-wise product and concatenation between the latent factors (see Figure 2)) can enhance the quality of venue recommendations. Such observations are consistent with the results reported by previous literature for NeuMF [10] and DRCF [1].

Next, we note that unlike the Brightkite and Foursquare checkin datasets, the Yelp dataset consists only of user-venue ratings, and hence the sequential properties of visits to venues are less likely to be observed. We observe that the RNN-based approaches (RNN and DREAM) that take the sequential properties of checkins into account are more effective than the traditional MF-based approaches (MF and BPR) across the Brightkite and Foursquare checkin datasets. However, both RNN and DREAM are less effective than BPR for the Yelp rating dataset because the sequential properties of rating data are less pronounced than the other LBSNs. This is likely due to users writing Yelp reviews after visiting the venues. In contrast, our proposed CRCF framework is still the most effective across the different types of datasets, which is indicative of the generalisability of CRCF. In addition, we observe that CARA, which incorporates the contextual information, is as effective as DRCF on the two checkin datasets in terms of the two used measures[12], while CARA outperforms DRCF on the Yelp dataset. These results demonstrate that contextual information plays an important role in enhancing the effectiveness of CAVR. By integrating CARA into CRCF, we can further enhance the quality of CAVR across three datasets in terms of HR@10 and NDCG@10.

line:s:factorsNext, Figure 3 reports the test performance of the CRCF framework and the baselines with respect to different dimensions of latent factors on the three datasets in terms of HR@10 and NDCG@10. From the figure, we observe that on the Brightkite dataset, the performances of DRCF, CARA and CRCF for both metrics increase as the dimensions of latent factors increase, while their performances decrease when the latent factors are set to 80. This effect can be clearly seen with CARA for the Brightkite and Foursquare datasets. These results are consistent with those reported in the previous literature (e.g. [10, 37]) where the performance of the

[12] Note that DRCF consists of three components (namely GRMF, MLRP and RMF models (see Section 3.2)), with each component having its own recurrent layer. Although CARA consists only one recurrent layer, it is as effective as DRCF. Moreover, in [2], we demonstrated that CARA significantly outperforms each individual component of DRCF.

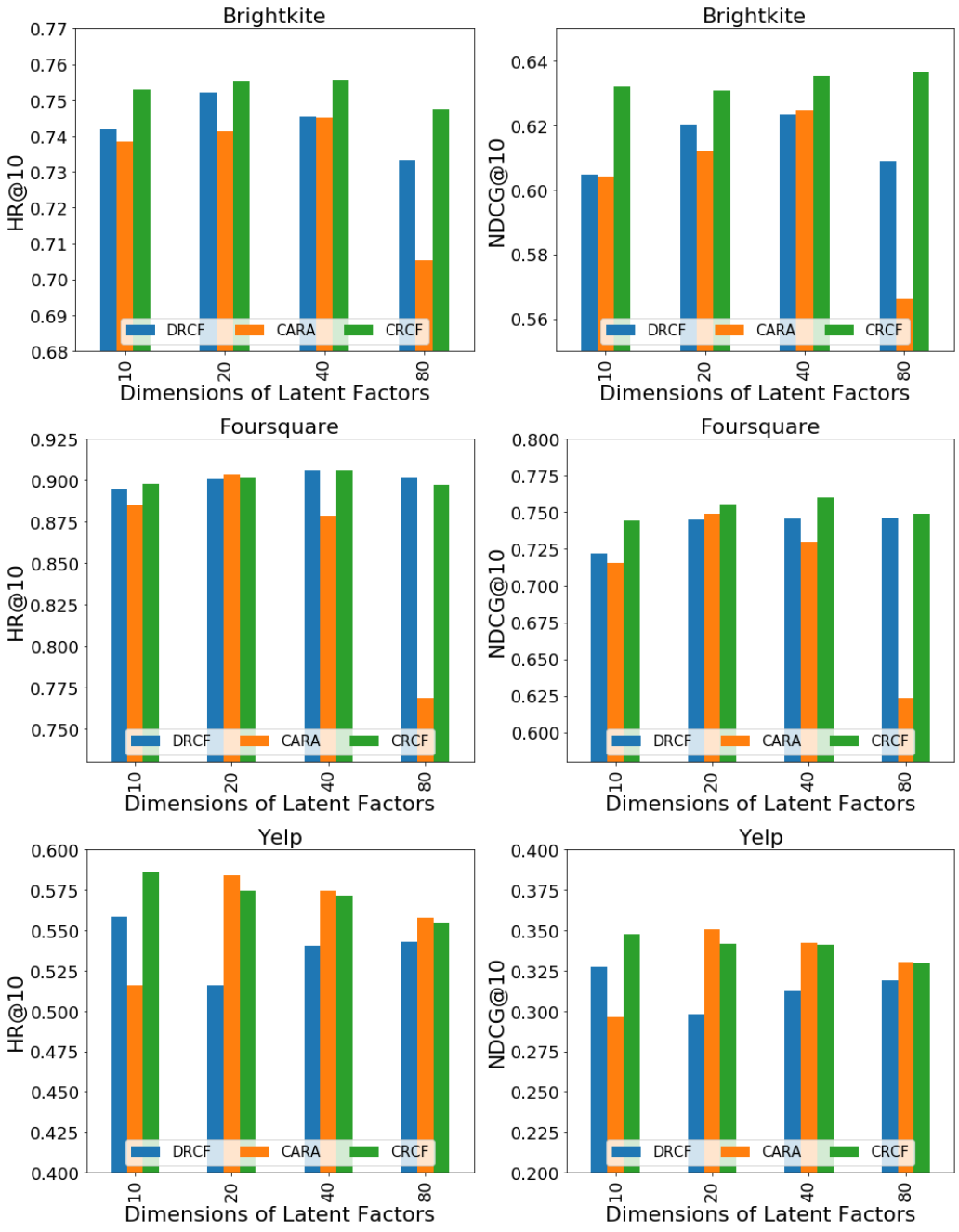


Figure 3: Test recommendation performances in terms of HR & NDCG of various approaches while varying the dimensions of latent factors. NB: The y-axis scale is adjusted for each plot to ensure legibility.

Deep Neural Network model decreases when its parameter size increases, due to overfitting. Interestingly, on the Yelp dataset, CARA outperforms CRCF when the dimensions of latent factors increase using both metrics. These results are intuitive because CARA has a lower number of parameters than CRCF (indeed, the number of parameters of CRCF is approximately the sum of CARA's and DRCF's parameters). Therefore, when we increase the number of latent factors in CARA, it can learn to more effectively capture the complex structure of user-venue interactions on the Yelp dataset. In contrast, CRCF already has a large number of parameters, increasing the dimensions of latent factors of CRCF is likely to degrade its effectiveness.line:e:factors

Within the second group of rows in Table 3, we further investigate the effectiveness of the CRCF framework by comparing with the baselines for the *Cold-Start Users*. Similar to the first group of rows in Table 3, the results in the second group demonstrate that CRCF consistently and significantly outperforms all baselines across the Brightkite and Yelp datasets on both measures. In particular, comparing the effectiveness of CRCF for cold-start users with DRCF and CARA, CRCF obtains 5.81-10% and 3.69-15.49% improvements in terms of NDCG, for the Brightkite and Yelp datasets, respectively. Although the performance of CRCF in alleviating the cold-start user problem is statistically indistinguishable from CARA and DRCF for the Foursquare dataset in terms of HR@10, CRCF significantly outperforms CARA in terms of NDCG@10 by 7%. This result suggests that the element-wise product and the concatenation of the latent factors used by CRCF play a more important role than the dot product of the latent factors used by CARA in generating more effective top-K venue recommendations for cold-start users. Overall, the results reported in the second group of rows in Table 3 demonstrate that our proposed CRCF framework is more effective than the DRCF framework and the CARA architecture in alleviating the cold-start user problem. Overall, in response to research question RQ1(a), we find that our proposed CRCF framework, which leverages the sequences of users' checkins as well as the contexts associated with the checkins, is effective for CAVR for both normal and cold-start users.

line:s:sec2

6.1.2. Usefulness of Dynamic Geo-based Negative Sampling

In this section, to address research question RQ2(a), we evaluate the usefulness of the *dynamic* geo-based negative sampling approach (Equation (3), denoted with the suffix $_{dgeo}$) in enhancing the robustness of var-

Table 4: As per Table 3. Performances in terms of HR@10 and NDCG@10 between various approaches that apply the dynamic geo-based negative sampling approach proposed in [1], denoted as *dgeo*

Normal Users Experiments						
	Brightkite		Foursquare		Yelp	
Model	HR	NDCG	HR	NDCG	HR	NDCG
MostPop	0.1462*	0.1010*	0.2009*	0.1167*	0.0739*	0.0334*
MostVisit	0.4032*	0.3473*	0.4733*	0.4290*	0.1083*	0.0528*
RecentVisit	0.4809*	0.4370*	0.4584*	0.4037*	0.1096*	0.0542*
GeoBPR	0.7339*	0.4672*	0.8216*	0.5395-	0.5570*	0.3032*
DRCF	0.7419*	0.6048*	0.8952*	0.7223*	0.5162*	0.2963*
CARA	0.7385*	0.6040*	0.8851*	0.7154*	0.5587*	0.3272*
CRCF	0.7528	0.6319	0.8981	0.7442	0.5861	0.3479
DRCF$_{dgeo}$	0.7852*	0.6210*	0.9095*	0.7214*	0.5618*	0.3064*
CARA$_{dgeo}$	0.7717*	0.6266*	0.9129*	0.7567*	0.6107*	0.3665*
CRCF$_{dgeo}$	**0.8029**	**0.6606**	**0.9260**	**0.7788**	**0.6548**	**0.3927**
Cold-Start Users Experiments						
	Brightkite		Foursquare		Yelp	
Model	HR	NDCG	HR	NCDG	HR	NDCG
MostPop	0.1155*	0.0778*	0.0584*	0.0286*	0.0714*	0.0316*
MostVisit	0.4285*	0.3789*	0.3506*	0.3175*	0.1044*	0.0489*
RecentVisit	0.4995*	0.4585*	0.3831*	0.3446*	0.1052*	0.0497*
GeoBPR	0.8093*	0.5262*	0.8312-	0.5486*	0.5802*	0.3202*
DRCF	0.7526*	0.5980*	0.8377*	0.6645*	0.5330*	0.3136*
CARA	0.7648*	0.6220*	0.8636*	0.6505*	0.5748*	0.3493*
CRCF	0.7782*	0.6582*	0.8571*	0.6967*	0.5913*	0.3622*
DRCF$_{dgeo}$	0.8094*	0.6199*	**0.8896**	0.7074	0.5877*	0.3318*
CARA$_{dgeo}$	0.8153*	0.6556*	0.8766	**0.7225**	0.6332*	0.3893*
CRCF$_{dgeo}$	**0.8557**	**0.6995**	0.8701	0.7152	**0.6612**	**0.4053**

ious approaches that used the traditional BPR negative sampling approach (Equation (2). In the first group of rows in Table 4, we report the effectiveness for *Normal Users* of the CRCF framework by comparing it with the state-of-the-art DRCF framework and the CARA architecture when incorporating the dynamic geo-based negative sampling approach. First, we observe similar results to those reported in [1], namely that the negative sampling approach can significantly improve the effectiveness of DRCF, CARA and CRCF in terms of HR@10 and NDCG@10 across the three datasets. For example, CRCF$_{dgeo}$ obtains over 6.65%, 3.1% and 11.72% improvements over CRCF in terms of HR@10 on the Brightkite, Foursquare and Yelp datasets, respectively. Note that DRCF$_{dgeo}$ and CARA$_{dgeo}$ also obtain similar percentage improvements over DRCF and CARA, respectively, across the three datasets. In addition, CRCF$_{dgeo}$ consistently and significantly outperforms all baselines that consider the geographical location of venues during the negative sampling process (i.e. GeoBPR, DRCF$_{dgeo}$ and CARA$_{dgeo}$) across all three datasets. These improvements and observed results demonstrate

that the *dynamic* geo-based negative sampling approach plays a crucial role in enhancing the effectiveness of DNN-based approaches. In addition, Figure 4 reports the test performance of CRCF$_{dgeo}$ and the baselines for each of the three datasets with all users over each training iteration. From the figure, we observe that CRCF$_{dgeo}$ outperforms all the baselines at every iteration and converges faster than others across the three datasets. Moreover, we observe that both DRCF$_{dgeo}$ and CARA$_{dgeo}$ are more effective that DRCF and CARA. However, on the Yelp dataset, we find that CRCF, which relies on the traditional BPR negative sampling approach [21] is more effective than DRCF$_{dgeo}$ at every iteration in terms of HR@10 and NDCG@10. These results demonstrate that the users' context plays an important role in enhancing the quality of CAVR. Indeed, the *dynamic* geo-based negative sampling approach may not be useful when the sequential properties of the users' observed feedback are less likely to be observed, as in the Yelp rating dataset. Hence, DRCF$_{dgeo}$ is less effective than CRCF for both measures on the Yelp dataset.

Next, within the second group of rows in Table 4, we further investigate the effectiveness of CRCF$_{dgeo}$, DRCF$_{dgeo}$ and CARA$_{dgeo}$, which all rely on the *dynamic* geo-based negative sampling approach, in the *Cold-Start Users* experiments. First, similar to the *Normal Users* experiments, we observe that the dynamic geo-based negative sampling approach can significantly improve the effectiveness of DRCF, CARA and CRCF in terms of HR@10 and NDCG@10 across the three datasets in the *Cold-Start Users* experiments. In particular, the results demonstrate that CRCF$_{dgeo}$ consistently and significantly outperforms all baselines across both the Brightkite and Yelp datasets on both measures. In particular, comparing the effectiveness of alleviating the cold-start users of CRCF$_{dgeo}$ with DRCF$_{dgeo}$ and CARA$_{dgeo}$, CRCF obtains approximately 5%, 4.42 - 12.5% improvements in terms of HR@10 for the Brightkite and Yelp datasets, respectively. Although the effectiveness of CRCF$_{dgeo}$ for the cold-start users is less than that of DRCF$_{dgeo}$ and CARA$_{dgeo}$ for the Foursquare dataset, there is no significant difference between CRCF$_{dgeo}$, DRCF$_{dgeo}$ and CARA$_{dgeo}$ in terms of HR and NDCG on the Foursquare dataset. These results demonstrate that the *dynamic* geo-based negative sampling approach can enhance the effectiveness of CRCF, DRCF and CARA in generating effective CAVR for the cold-start users.

We further investigate the usefulness of the dynamic geo-based negative sampling approach and the CARA architecture in enhancing the effectiveness of CRCF$_{dgeo}$ under different settings for CAVR. Note that CARA leverages

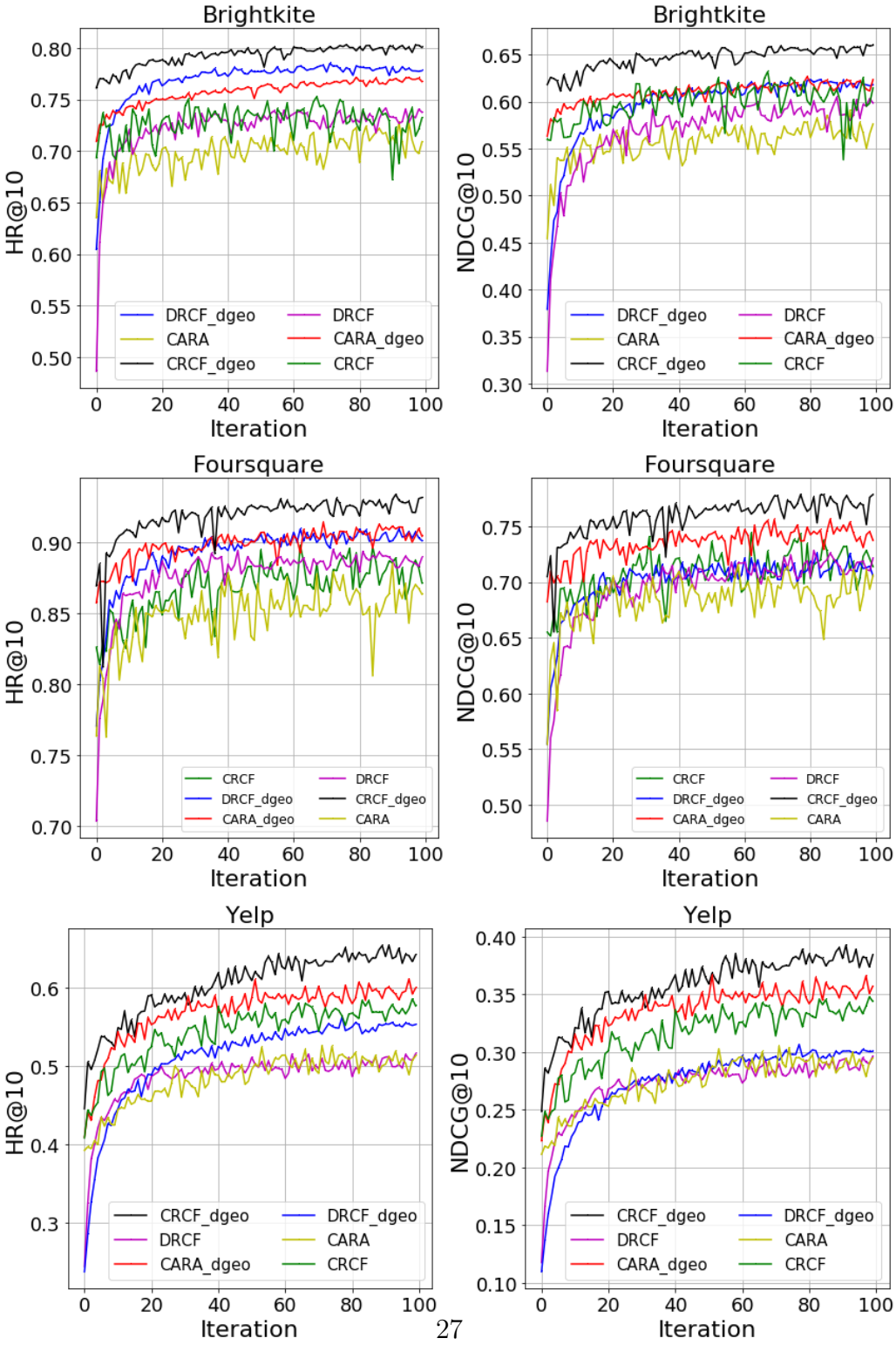
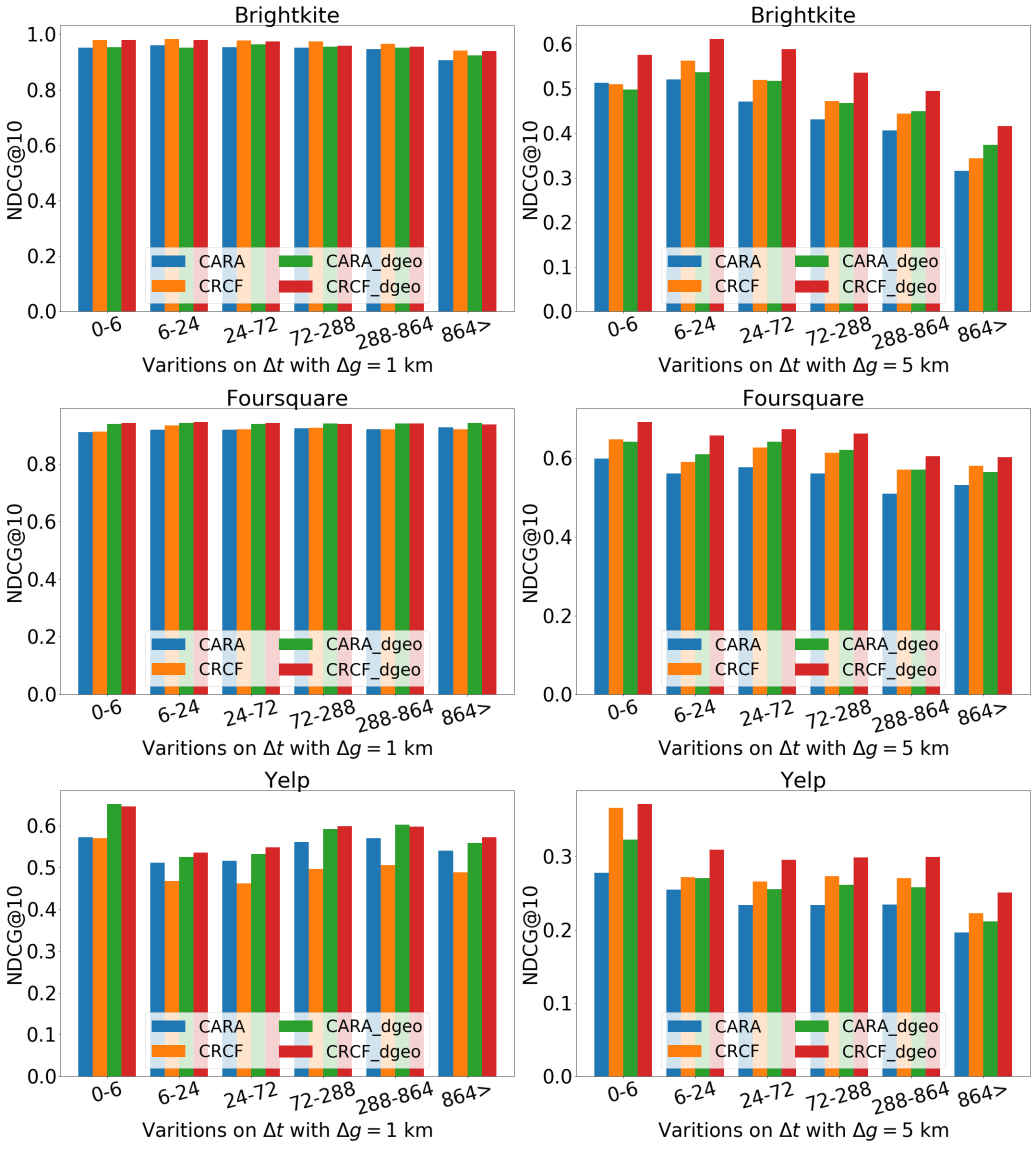


Figure 4: Test recommendation performances in terms of HR & NDCG of various approaches with respect to the number of iterations.

Figure 5: Performance of various approaches in terms of NDCG@10 on the Brightkite, Foursquare and Yelp datasets by varying the time interval Δt in terms of hours with the fixed values of the geographical distances Δg (1 and 5 km).



the time interval and geographical distance between two successive checkins to model the user's dynamic preferences, motivating the integration of CARA into our proposed CRCF framework (as described in Section 4). In particular, Figure 5 presents the performances on the Brightkite, Foursquare and Yelp datasets – in terms of NDCG@10 – of various approaches, by considering the users with particular time intervals Δt (hours) and geographical distances Δg (km) between their last checkin and the ground-truth checkin. For example, if a user checks in at venues A, B and C in a sequence, his/her last checkin is at venue B and his/her ground-truth checkin is at venue C. Then, we calculate the distance Δg and the time interval Δt between venues B and C. Note that these two checkins may occur at the same venue, hence the distance $\Delta g = 0$, while the time interval Δt between these two checkins is such that $\Delta t > 0$. First, the results from Figure 5 demonstrate that CRCF consistently outperforms CARA across the three datasets in terms of NDCG@10 on various time intervals Δt and geographical distances Δg. These results suggest that the neural architecture (i.e. the element-wise and concatenation operations of latent factors, described in Section 4) in CRCF can effectively integrate CARA, hence obtaining the improvements over CARA on both settings. Moreover, the experimental results using a fixed geographical distance of $\Delta g = 5$ km on the right-hand plots in Figure 5 demonstrate that the effectiveness of all approaches on the three datasets decreases as the time intervals between two successive checkins increase. These results suggest that users are less likely to be influenced by distant venues they visited a long time ago, which are consistent with results previously reported in the literature [2]. In contrast, the performances of all approaches on a fixed geographical distance of $\Delta g = 1$ km setting are relatively stable on the Brightkite and Foursquare datasets. Intuitively, nearby venues visited by users are more likely to influence the users' preferences for their next venues regardless of when those nearby venues were visited. As mentioned above, the sequential properties are less likely to be observed from the user-venue rating Yelp dataset. Hence, unlike the Brightkite and Foursquare checkin datasets, the *dynamic* geo-based negative sampling approach may not be useful in enhancing the performances of DRCF, CARA and CRCF on the Yelp dataset. Furthermore, comparing the approaches that apply the dynamic geo-based negative sampling approach, we find that the effectiveness of both CARA_{dgeo} and CRCF_{dgeo} across the three datasets on different settings can be enhanced by the dynamic negative sampling approach. In particular, CRCF_{dgeo} is the most effective approach compared to all baselines across the three datasets on various settings. Overall, in

response to research question RQ2(a), we find that the dynamic geo-based negative sampling approach can effectively improve the performances of the CRCF framework for CAVR on various settings that consider different time intervals and geographical distances between users' two successive checkins.

6.2. Robustness Evaluation

In this section, we evaluate the robustness of the CRCF and DRCF frameworks as well as the CARA architecture using the risk-sensitive measures (i.e. Reward & Risk and U_{risk}), proposed by Wang *et al.* [24] to quantify any underperformance of DRCF, CARA and CRCF compared to the BPR model (Section 6.2.1).[13] Apart from the risk-sensitive measures, we also use the T_{risk} measure, proposed by Dinccer *et al.* [25], to evaluate whether a given framework or model exhibits a significant risk compared to the BPR model. In particular, we test the significance of an observed risk-reward tradeoff score between a target model and the BPR model by using T_{risk} as the test statistic of the Student's t-test for matched pairs. In addition, we evaluate the usefulness of the *dynamic* geo-based negative sampling approach in enhancing the robustness of the CRCF framework (Section 6.2.2).

6.2.1. Robustness of the CRCF framework

Tables 5 & 6 report the robustness of the CRCF framework in comparison with the DRCF framework and the CARA architecture – which do not apply the dynamic geo-based negative sampling approach – on the three datasets in terms of different measures under the *Normal Users* and *Cold-Start Users* experiments, respectively. For instance, the Wins/Losses row shows the ratio of the number of users that benefit or do not benefit from a particular model compared to the BPR model. On analysing Table 5, in terms of the robustness of the approaches, we find that CRCF is the most robust framework by consistently having the lowest Risk/Losses and the highest Reward/Wins in comparison with DRCF and CARA across the three datasets. In particular, CRCF can generate a more effective ranked list of venues than BPR (i.e. NDCG@10 is improved by CRCF compared to BPR) for 45.48%, 54.56% and 33.32% of users on the Brightkite, Foursquare and Yelp datasets, respectively. CRCF performs less effectively than BPR (i.e. NDCG@10 is degraded

[13] Indeed, we argue that BPR is a widely used baseline in recommendation systems, which is akin to the use of BM25 in web search, and hence is appropriate as our robust baseline for risk-sensitive evaluation.

Table 5: The robustness of various approaches in comparison with the BPR baseline in terms of NDCG@10 on three datasets for *Normal* users. T_{risk} scores greater than +2 or less than -2 indicate that a two-tailed paired t-test gives significance with $p < 0.05$. T_{risk} scores greater than +2 are indicated with $*$. The best score w.r.t. each risk-sensitive measure is highlighted in bold.

Dataset	Measure	DRCF	CARA	CRCF
Brightkite	Risk	0.052	0.049	**0.041**
	Reward	0.224	0.220	**0.240**
	Wins/Losses	6297/1805	6159/1687	**6538/1483**
	Wins%/Losses%	43.80/12.55	42.84/11.73	**45.48/10.31**
	U_{risk} $\alpha = 1$	0.119	0.122	0.157
	T_{risk} $\alpha = 1$	38.763*	40.390*	52.161*
	U_{risk} $\alpha = 5$	-0.090	-0.073	-0.007
	T_{risk} $\alpha = 5$	-24.066	-20.195	-2.051
Foursquare	Risk	**0.019**	0.029	0.022
	Reward	0.258	0.261	**0.283**
	Wins/Losses	5723/**644**	5450/910	**5876**/711
	Wins%/Losses%	53.14/**5.98**	50.61/8.45	**54.56**/6.60
	U_{risk} $\alpha = 1$	0.220	0.202	0.238
	T_{risk} $\alpha = 1$	71.063*	60.024*	71.698*
	U_{risk} $\alpha = 5$	0.142	0.084	0.147
	T_{risk} $\alpha = 5$	44.010*	22.915*	42.049*
Yelp	Risk	0.071	0.075	**0.070**
	Reward	0.097	0.133	**0.148**
	Wins/Losses	9232/7311	11820/7518	**12980/7064**
	Wins%/Losses%	23.70/18.77	30.35/19.30	**33.32/18.13**
	U_{risk} $\alpha = 1$	-0.045	-0.017	0.009
	T_{risk} $\alpha = 1$	-30.236	-10.322	5.025*
	U_{risk} $\alpha = 5$	-0.330	-0.320	-0.272
	T_{risk} $\alpha = 5$	-142.903	-126.541	-111.453

by CRCF compared to BPR) for 10.32%, 6.60% and 18.14% of users on the Brightkite, Foursquare and Yelp datasets, respectively. In addition, Figure 6 reports the wins-losses histogram of CRCF and the baselines on the three datasets. From the plots for the *Normal Users* experiments (left-hand plots of Figure 6), we observe that CRCF has consistently smaller changes in NDCG@10 on all bins on the left side of the vertical line and larger changes in NDCG@10 on the right side of the vertical line than the baselines across the three datasets. Moreover, we observe that, at $\alpha = 1$ (which emphasises risk twice over reward), the calculated U_{risk} scores of DRCF, CARA and CRCF are significantly higher than BPR, at $p < 0.05$ (as $T_{risk} > 2$), across the Brightkite and Foursquare datasets, while only the U_{risk} score of CRCF on the Yelp dataset exhibits significant risk ($T_{risk} < 2$). These results demonstrate that it is highly likely that CRCF will not perform worse than the BPR baseline across the three datasets, while both DRCF and CARA, having $T_{risk} < -2$ at $\alpha = 1$, may underperform on the Yelp dataset (i.e. perform worse than BPR). Overall, in response to research question RQ1(b), we find that our proposed CRCF framework is robust and less likely to perform

Table 6: The robustness of various approaches in comparison with the BPR baseline in terms of NDCG@10 on three datasets for the *Cold-Start* users. T_{risk} scores greater than +2 or less than -2 indicate that a two-tailed paired t-test gives significance with $p < 0.05$. T_{risk} scores greater than +2 are indicated with $*$. The best score w.r.t. each risk-sensitive measure is highlighted in bold.

Dataset	Measure	DRCF	CARA	CRCF
Brightkite	Risk	0.082	0.060	**0.047**
	Reward	0.190	0.192	**0.215**
	Wins/Losses	2183/1043	2221/803	**2394/661**
	Wins%/Losses%	39.13/18.69	39.81/14.39	**42.91/11.85**
	U_{risk} $\alpha = 1$	0.025	0.071	0.121
	T_{risk} $\alpha = 1$	4.871*	15.045*	25.832*
	U_{risk} $\alpha = 5$	-0.304	-0.170	-0.067
	T_{risk} $\alpha = 5$	-40.821	-27.581	-12.069
Foursquare	Risk	0.056	0.067	**0.041**
	Reward	0.224	0.221	**0.241**
	Wins/Losses	**73/20**	57/86	65/81
	Wins%/Losses%	**47.40/12.98**	37.01/55.84	42.20/52.59
	U_{risk} $\alpha = 1$	0.113	0.087	0.160
	T_{risk} $\alpha = 1$	3.785*	2.813*	5.498*
	U_{risk} $\alpha = 5$	-0.109	-0.181	-0.003
	T_{risk} $\alpha = 5$	-2.955	-4.438	-0.092
Yelp	Risk	0.079	0.079	**0.078**
	Reward	0.098	0.134	**0.145**
	Wins/Losses	1593/1421	2119/1374	**2266/1328**
	Wins%/Losses%	23.07/20.58	30.69/19.90	**32.82/19.23**
	U_{risk} $\alpha = 1$	-0.059	-0.024	-0.010
	T_{risk} $\alpha = 1$	-16.220	-5.982	-2.411
	U_{risk} $\alpha = 5$	-0.374	-0.341	-0.321
	T_{risk} $\alpha = 5$	-63.249	-55.016	-51.850

worse than the BPR baseline for CAVR for *Normal* users.

Next, Table 6 reports the risk measures for the *Cold-Start Users*, using the same notations as Table 5. In Table 6, we observe that CRCF has consistently lower Risk/Losses and higher Reward/Wins than DRCF and CARA across the Brightkite and Yelp datasets for the *Cold-Start Users*. For example, CRCF is more robust than DRCF and CARA in terms of Wins as it can generates more effective venue suggestions than BPR for 2,394 users on the Brightkite dataset, while DRCF and CARA can only generate more effective venue suggestions than BPR for 2,183 and 2,221 users, respectively. Similar results in terms of Wins for DRCF, CARA and CRCF can also be observed on the Yelp dataset. In addition, CRCF exhibits less risk than DRCF and CARA at generating less effective venue suggestions than BPR. For example, on the Brightkite dataset, CRCF only generates less effective venue suggestions than BPR for 661 users, while DRCF and CARA generate less effective venue suggestions than BPR for 1,043 and 803 users, respectively. We observe similar results for DRCF, CARA and CRCF in terms of Losses on the Yelp dataset.

Moreover, we observe that, at $\alpha = 1$, the U_{risk} scores of DRCF, CARA and CRCF exhibit significant improvements across the Brightkite and Foursquare datasets. These results demonstrate that there is no significant risk that these three approaches will perform worse than the BPR baseline for the *Cold-Start* users for the Brightkite and Foursquare datasets. However, these three approaches are likely to perform worse than BPR on the Yelp dataset when both $\alpha = 1$ and $\alpha = 5$. Figure 6 (left-hand plots) shows that CRCF consistently has a larger number of positive changes over BPR than DRCF and CARA on the Brightkite and Yelp datasets across all positive bins, and likewise has a smaller number of negative changes over BPR than DRCF and CARA across negative the bins. However, for the Foursquare dataset, CRCF has a higher Reward and lower Risk than DRCF on average, while the number of Losses of DRCF is lower than that of CRCF. Likewise, the number of Wins of DRCF is higher than CRCF. For example, there are only 13% of cold-start users (20 out of 154 cold-start users) on the Foursquare dataset whose recommendations generated by DRCF are less effective than BPR, while 52% of cold-start users (81 out of 154 cold-start users) on Foursquare get less effective recommendations from CRCF than BPR. These results can be clearly observed in Figure 6 on the Foursquare dataset for the *Cold-Start Users* experiments where CARA and CRCF obtain a large number of negative changes in terms of NDCG@10 over BPR for $0 < \text{NDCG} \leq 0.2$. Overall, in response to research question RQ1(b), we find that our proposed CRCF framework is robust and less likely to perform worse than the BPR baseline for CAVR for the *Cold-Start* users.

We further investigate the robustness of the CRCF framework in comparison with the DRCF framework and the CARA architecture using the T_{risk} score. Figure 7 demonstrates the change in the T_{risk} scores of the approaches for various risk-sensitivity α parameter values from 0 to 15 under the *Normal* and *Cold-Start Users* experiments. Note that, as mentioned in Section 5, the risk-sensitivity α parameter controls the risk-reward tradeoff of the U_{risk} and T_{risk} scores. Indeed, as α increases, the tradeoff between risk and reward for each model changes in favour of risk compared to reward. T_{risk} scores greater than +2 (indicated by the red horizontal line in the figure) or less than 2 (indicated by the blue-dashed horizontal line in the figure) exhibit significant differences from the baseline according to a two-tailed paired t-test with $p < 0.05$. On analysing the left-hand plots in Figure 7, with respect to the *Normal* experiments, at $\alpha = 1$, we observe that all approaches (DRCF, CARA and CRCF) are significantly less risky than

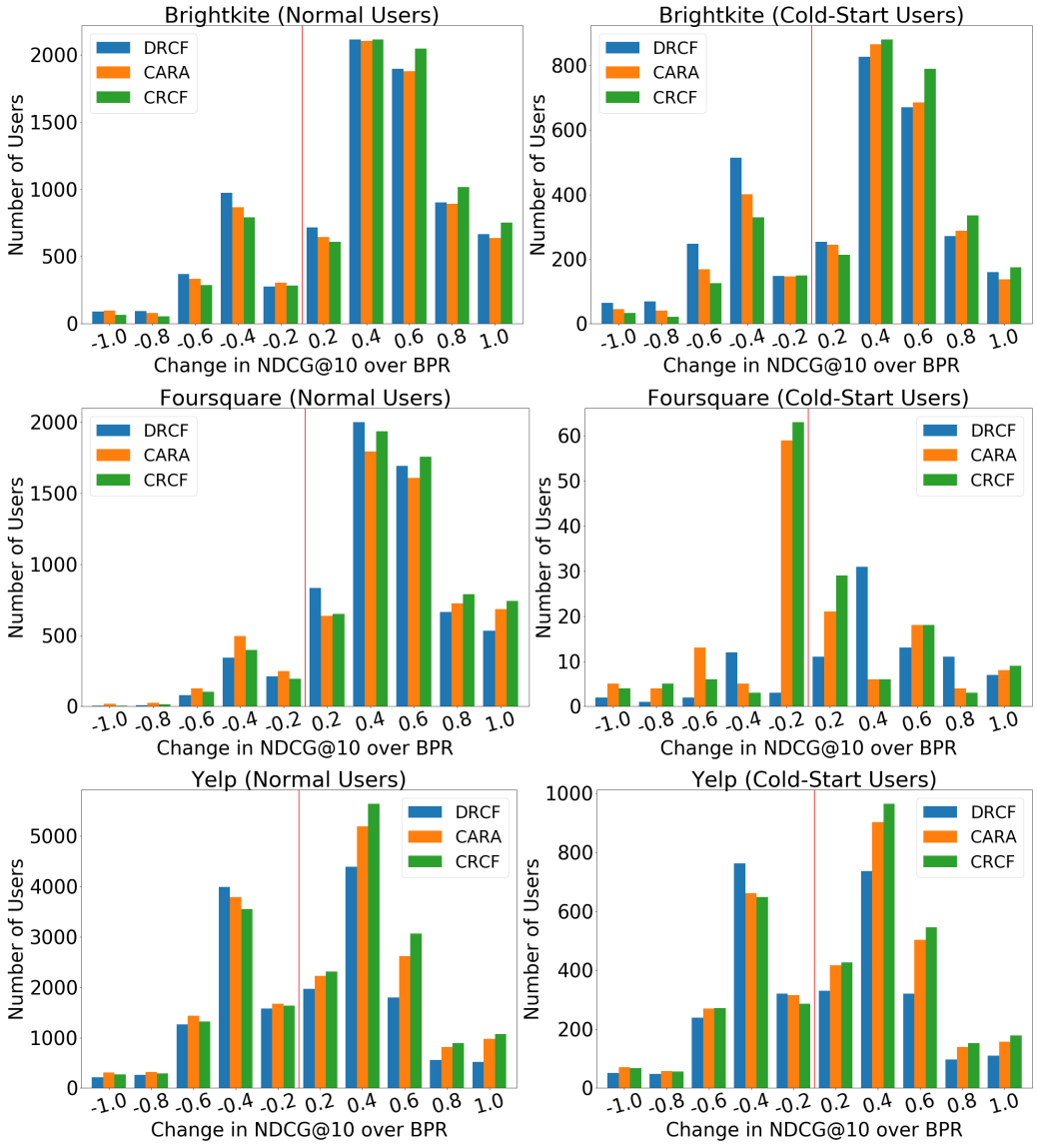


Figure 6: Wins-Losses histograms of new models (DRCF, CARA and CRCF) in comparison with the BPR ranking baseline model under the *Normal* and *Cold-start Users* settings. The vertical line in the figures separates the wins and losses histograms. We omit to report the number of users with no changes in NDCG@10 over BPR (i.e. the target model is as effective as BPR).

BPR across the three datasets. Moreover, we observe that as α increases, on the Brightkite and Yelp datasets, CRCF is significantly less risky than BPR when $\alpha = 4$ and $\alpha = 1$, respectively, while DRCF and CARA are not. On the Foursquare dataset, CRCF and DRCF are significantly less risky than BPR until $\alpha = 11$, while CARA is significantly less risky than BPR until $\alpha = 7$.

Next, on analysing the right-hand plots in Figure 7, regarding the robustness of DRCF, CARA and CRCF under the *Cold-Start* experiments, we observe that, at $\alpha = 1$, all approaches are significantly less risky than BPR. However, as α increases to 3, CRCF is the only approach that is less risky than the BPR baseline across the Brightkite and Foursquare datasets. Moreover, comparing CRCF with either DRCF or CARA, we observe that CRCF is only significantly less risky than DRCF when $\alpha = 1$ on the Brightkite dataset for the *Cold-Start* users. Overall, in response to research question RQ1(b), we find that our proposed CRCF framework is less risky for deployment to users, in that it only exhibits real risk compared to BPR for higher values of α than the existing state-of-the-art, DRCF and CARA approaches, for both *Normal* and *Cold-Start* users experiments.

6.2.2. Usefulness of Dynamic Geo-based Negative Sampling for Robustness

In this section, in addressing research question RQ2(b), we evaluate the usefulness of the *dynamic* geo-based negative sampling in improving the robustness of the CRCF framework for the *Normal Users* experiments. In Table 7, we first observe that the *dynamic* geo-based negative sampling approach can consistently enhance the robustness of the CRCF framework across the three datasets. In particular, in comparison with DRCF_{dgeo} and CARA_{dgeo}, CRCF_{dgeo} is the most robust framework, as it generates more effective venue suggestions than BPR for 49.94% , 59.16% and 40.55% of users on the Brightkite, Foursquare and Yelp datasets, respectively. Moreover, comparing CRCF_{dgeo} and CRCF, we observe that the dynamic geo-based negative sampling approach can enhance the Reward score of CRCF by approximately 4-7% and can reduce the Risk score of CRCF by approximately 0.6-3%. In addition, comparing the T_{risk} scores of CRCF and CRCF_{dgeo} on the Brightkite dataset, at $\alpha = 5$, we observe that CRCF_{dgeo} is less likely to exhibit a real risk of performing worse than the BPR baseline, while CRCF is not. In addition, Figure 8 reports the robustness of CRCF and CRCF_{dgeo} on the three datasets. From the left-hand plots in Figure 8 on the *Normal Users* experiments, we observe that CRCF_{dgeo} has consistently lower changes

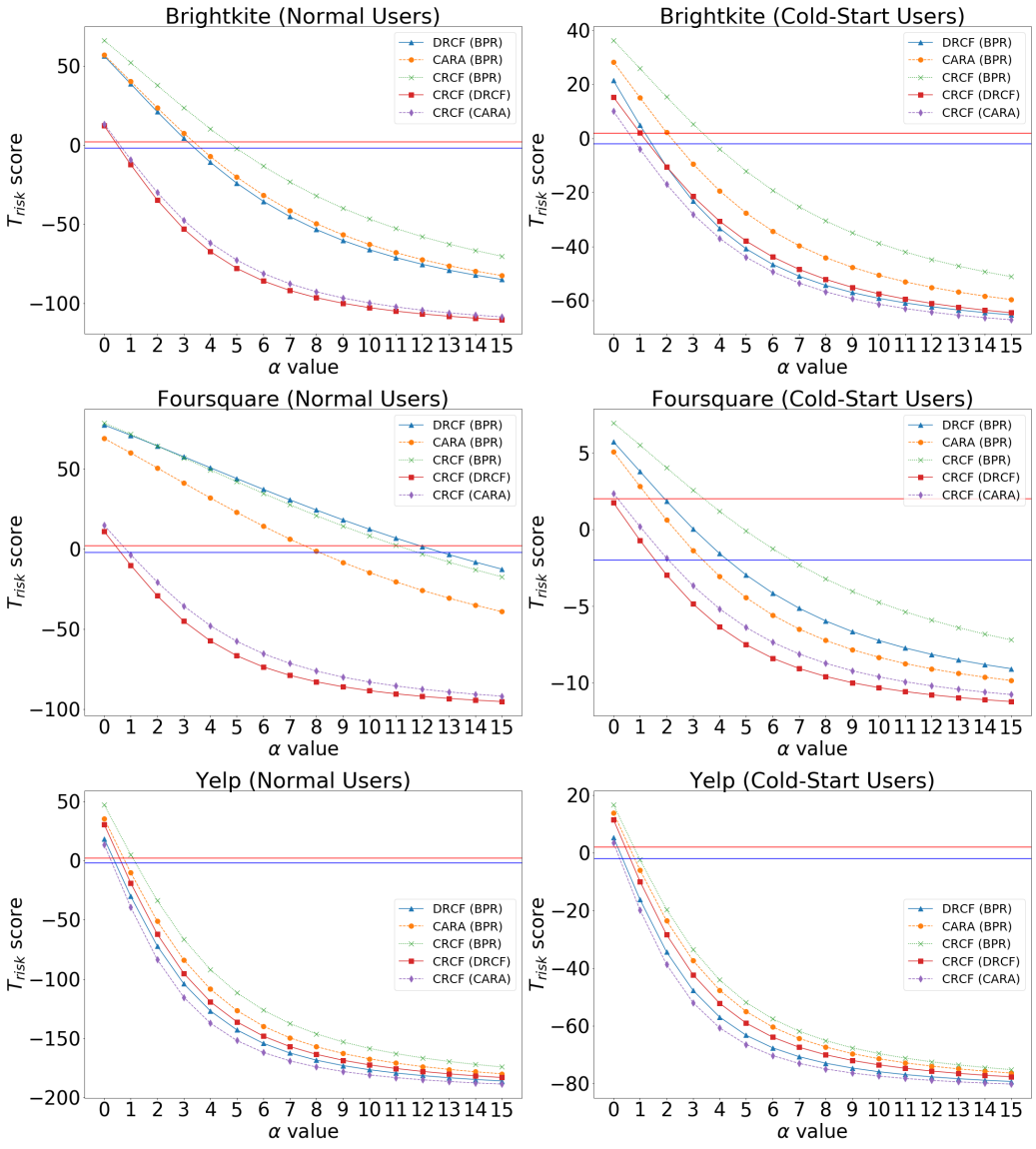


Figure 7: The change in standardised T_{risk} scores for DRCF, CARA and CRCF with respect to the BPR model, denoted inside the parentheses, over different α values under the *Normal* and *Cold-Start Users* experiments.

Table 7: The robustness of various approaches that incorporate the *dynamic* geo-based negative sampling in comparison with the BPR baseline in terms of NDCG@10 on three datasets for *Normal* users. T_{risk} scores greater than +2 or less than -2 indicate that a two-tailed paired t-test gives significance with $p < 0.05$. T_{risk} scores greater than +2 are indicated with $*$. The best score w.r.t. each risk-sensitive measure is highlighted in bold.

Dataset	Measure	DRCF	CARA	CRCF	$DRCF_{dgeo}$	$CARA_{dgeo}$	$CRCF_{dgeo}$
Brightkite	Risk	0.052	0.049	0.041	0.049	0.044	**0.031**
	Reward	0.224	0.220	0.240	0.237	0.238	**0.258**
	Wins/Losses	6297/1805	6159/1687	6538/1483	6836/1677	6660/1568	**7179/1139**
	Wins%/Losses%	43.80/12.55	42.84/11.73	45.48/10.31	47.55/11.66	46.33/10.90	**49.94/7.92**
	U_{risk} $\alpha = 1$	0.119	0.122	0.157	0.138	0.149	0.196
	T_{risk} $\alpha = 1$	38.763*	40.390*	52.161*	45.453*	49.212*	67.637*
	U_{risk} $\alpha = 5$	-0.090	-0.073	-0.007	-0.059	-0.029	0.073
	T_{risk} $\alpha = 5$	-24.066	-20.195	-2.051	-16.113	-8.170	23.171*
Foursquare	Risk	**0.019**	0.029	0.022	0.022	0.027	0.021
	Reward	0.258	0.261	0.283	0.260	0.301	**0.316**
	Wins/Losses	5723/**644**	5450/910	5876/711	5798/749	6153/821	**6371**/652
	Wins%/Losses%	43.80/12.55	42.84/11.73	45.48/10.31	53.84/6.95	57.14/7.62	**59.16/6.05**
	U_{risk} $\alpha = 1$	0.220	0.202	0.23	0.215	0.245	0.274
	T_{risk} $\alpha = 1$	71.063*	60.024*	71.698*	68.998*	70.814*	80.755*
	U_{risk} $\alpha = 5$	0.142	0.084	0.147	0.125	0.134	0.189
	T_{risk} $\alpha = 5$	44.010*	22.915*	42.049*	37.960*	36.110*	53.524*
Yelp	Risk	0.071	0.075	0.070	0.072	0.067	**0.059**
	Reward	0.097	0.133	0.148	0.109	0.165	**0.183**
	Wins/Losses	9232/7311	11820/7518	12980/7064	10927/7184	14270/6721	**15807/6117**
	Wins%/Losses%	23.70/18.77	30.35/19.30	33.32/18.13	28.05/18.44	36.64/17.25	**40.58/15.70**
	U_{risk} $\alpha = 1$	-0.045	-0.017	0.009	-0.035	0.029	0.064
	T_{risk} $\alpha = 1$	-30.236	-10.322	5.025*	-22.989	16.783*	36.938*
	U_{risk} $\alpha = 5$	-0.330	-0.320	-0.272	-0.323	-0.243	-0.174
	T_{risk} $\alpha = 5$	-142.903	-126.541	-111.453	-137.942	-99.610	-76.218

in NDCG@10 on all negative bins (i.e. to the left side of the vertical line) and higher changes in NDCG@10 on all positive bins (i.e. the right side of the vertical line) than CRCF across the three datasets. Furthermore, Figure 9 reports the wins-losses histograms of $DRCF_{dgeo}$, $CARA_{dgeo}$ and $CRCF_{dgeo}$ on the three datasets. From the left-hand plots in Figure 9, on the *Normal Users* experiments, we observe that $CRCF_{dgeo}$ consistently has lower changes in NDCG@10 on all negative bins (i.e. to the left side of the vertical line) and higher changes in NDCG@10 on all positive bins (i.e. to the right side of the vertical line) than $DRCF_{dgeo}$ and $CARA_{dgeo}$ across the three datasets.

Next, we evaluate the usefulness of the *dynamic* geo-based negative sampling approach in improving the robustness of the CRCF framework for the *Cold-Start Users* experiments. Similarly to the results reported in Table 7, in Table 8, we find that the *dynamic* geo-based negative sampling approach can consistently improve the robustness of CRCF on the Brightkite and Yelp datasets for the *Cold-Start Users* experiments. For example, $CRCF_{dgeo}$ obtains approximately 7% and 3% improvements in the Reward and Risk scores over DRCF on the Brightkite and Yelp datasets, respectively. Moreover, com-

Table 8: The robustness of various approaches that incorporate the *dynamic* geo-based negative sampling in comparison with the BPR baseline in terms of NDCG@10 on three datasets for the *Cold-Start* users. T_{risk} scores greater than +2 or less than -2 indicate that a two-tailed paired t-test gives significance with $p < 0.05$. T_{risk} scores greater than +2 are indicated with $*$. The best score w.r.t. each risk-sensitive measure is highlighted in bold.

Dataset	Measure	DRCF	CARA	CRCF	$DRCF_{dgeo}$	$CARA_{dgeo}$	$CRCF_{dgeo}$
Brightkite	Risk	0.082	0.060	0.047	0.078	0.051	**0.033**
	Reward	0.190	0.192	0.215	0.207	0.216	**0.241**
	Wins/Losses	2183/1043	2221/803	2394/661	2418/974	2489/706	**2767/486**
	Wins%/Losses%	39.13/18.69	39.81/14.39	42.91/11.85	43.34/11.66	44.62/10.90	**49.60/7.92**
	U_{risk} $\alpha = 1$	0.025	0.071	0.121	0.051	0.114	0.176
	T_{risk} $\alpha = 1$	4.871*	15.045*	25.832*	9.871*	24.273*	39.980*
	U_{risk} $\alpha = 5$	-0.304	-0.170	-0.067	-0.262	-0.090	0.046
	T_{risk} $\alpha = 5$	-40.821	-27.581	-12.069	-35.896	-15.663	9.396*
Foursquare	Risk	0.056	0.067	**0.041**	0.043	0.049	0.049
	Reward	0.224	0.221	0.241	0.254	**0.275**	0.268
	Wins/Losses	**73/20**	57/86	65/81	**71**/75	**73**/73	70/78
	Wins%/Losses%	**47.40/12.98**	37.01/55.84	42.20/52.59	46.10/48.70	**47.40**/47.40	45.45/50.64
	U_{risk} $\alpha = 1$	0.113	0.087	0.160	0.168	0.178	0.170
	T_{risk} $\alpha = 1$	3.785*	2.813*	5.498*	5.825*	5.796*	5.536*
	U_{risk} $\alpha = 5$	-0.109	-0.181	-0.003	-0.004	-0.018	-0.024
	T_{risk} $\alpha = 5$	-2.955	-4.438	-0.092	-0.115	-0.491	-0.674
Yelp	Risk	0.079	0.079	0.078	0.075	0.072	**0.066**
	Reward	0.098	0.134	0.145	0.113	0.166	**0.177**
	Wins/Losses	1593/1421	2119/1374	2266/1328	1962/1320	2548/1257	**2724/1195**
	Wins%/Losses%	23.07/20.58	30.69/19.90	32.82/19.23	28.42/19.12	36.91/18.20	**39.46/17.31**
	U_{risk} $\alpha = 1$	-0.059	-0.024	-0.010	-0.038	0.023	0.045
	T_{risk} $\alpha = 1$	-16.220	-5.982	-2.411	-10.301	5.468*	10.617*
	U_{risk} $\alpha = 5$	-0.374	-0.341	-0.321	-0.339	-0.264	-0.220
	T_{risk} $\alpha = 5$	-63.249	-55.016	-51.850	-58.746	-44.224	-38.381

paring the T_{risk} scores of CRCF and $CRCF_{dgeo}$ on the Brightkite dataset, at $\alpha = 5$, we observe that $CRCF_{dgeo}$ is less likely to exhibit a real risk of performing worse than the BPR baseline, while CRCF is not. Similarly, at $\alpha = 1$, on the Yelp dataset, we find that CRCF is likely to be under a real risk of performing worse than the BPR baseline, while $CRCF_{dgeo}$ is not. Next, the right-hand plots in Figure 8 report that $CRCF_{dgeo}$ has consistently a larger number of positive changes in NDCG@10 over BPR across all positive bins in comparison with CRCF across the three datasets. These significant improvements in the T_{risk} scores of $CRCF_{dgeo}$ compared to CRCF demonstrate that the *dynamic* geo-based negative sampling approach can significantly reduce the risk of the CRCF framework in performing worse than the BPR baselines. Furthermore, on analysing the right-hand plots in Figure 9, we observe that $CRCF_{dgeo}$ has consistently a larger number of positive changes in NDCG@10 over BPR across all positive bins in comparison with $DRCF_{dgeo}$ and $CARA_{dgeo}$ on the Brightkite and Yelp datasets. Overall, in response to research question RQ2(b), we find that the *dynamic* geo-based negative

sampling approach can significantly reduce the risk of our proposed CRCF framework in performing worse than the BPR baseline for both normal and cold-start users experiments.

Next, we evaluate the usefulness of the *dynamic* geo-based negative sampling approach in enhancing the robustness of the CRCF framework using the T_{risk} score. Similar to Figure 7, Figure 10 demonstrates the change in the T_{risk} scores of the various approaches with the *dynamic* geo-based negative under *Normal* and *Cold-Start Users* experiments. With respect to the *Normal* experiments, as α increases, we observe that $CRCF_{dgeo}$ is significantly less risky than the BPR baseline until $\alpha = 7, \alpha = 12$ and $\alpha = 2$ for the Brightkite, Foursquare and Yelp datasets, respectively, while CRCF is not. These results suggest that the *dynamic* geo-based negative sampling approach can significantly improve the robustness of our proposed CRCF framework (i.e. reducing the chance generating less effective recommendations than BPR). Overall, in response to research question RQ2(b), we observe further evidence that the *dynamic* geo-based negative sampling approach reduces the risk of our proposed CRCF framework, for both *Normal* and *Cold-Start* users.

7. Conclusions

In this article, we proposed a novel Contextual Recurrent Collaborative Filtering framework (CRCF) for Context-Aware Venue Recommendation (CAVR). Our proposed framework is built on top of two state-of-the-art deep neural network recommendation approaches, namely the Deep Recurrent Collaborative Filtering (DRCF) framework and the Contextual Attention Recurrent Architecture (CARA). By exploiting both DRCF and CARA, CRCF can effectively capture the complex structure of the users' dynamic preferences by considering their preferred context (i.e. time of the day) as well as the contextual information associated with the sequence of user's checkins.

Our comprehensive experiments on three large-scale datasets from Brightkite, Foursquare and Yelp demonstrated the significant improvements of our proposed CRCF framework for CAVR in comparison with various existing state-of-the-art venue recommendation approaches in both normal and cold-start settings. Moreover, our experimental results showed that CRCF is more robust than the baseline approaches in both the normal and cold-start settings. Indeed, the CRCF framework exhibited significantly less risk than both the DRCF framework and the CARA architecture across the three used datasets.

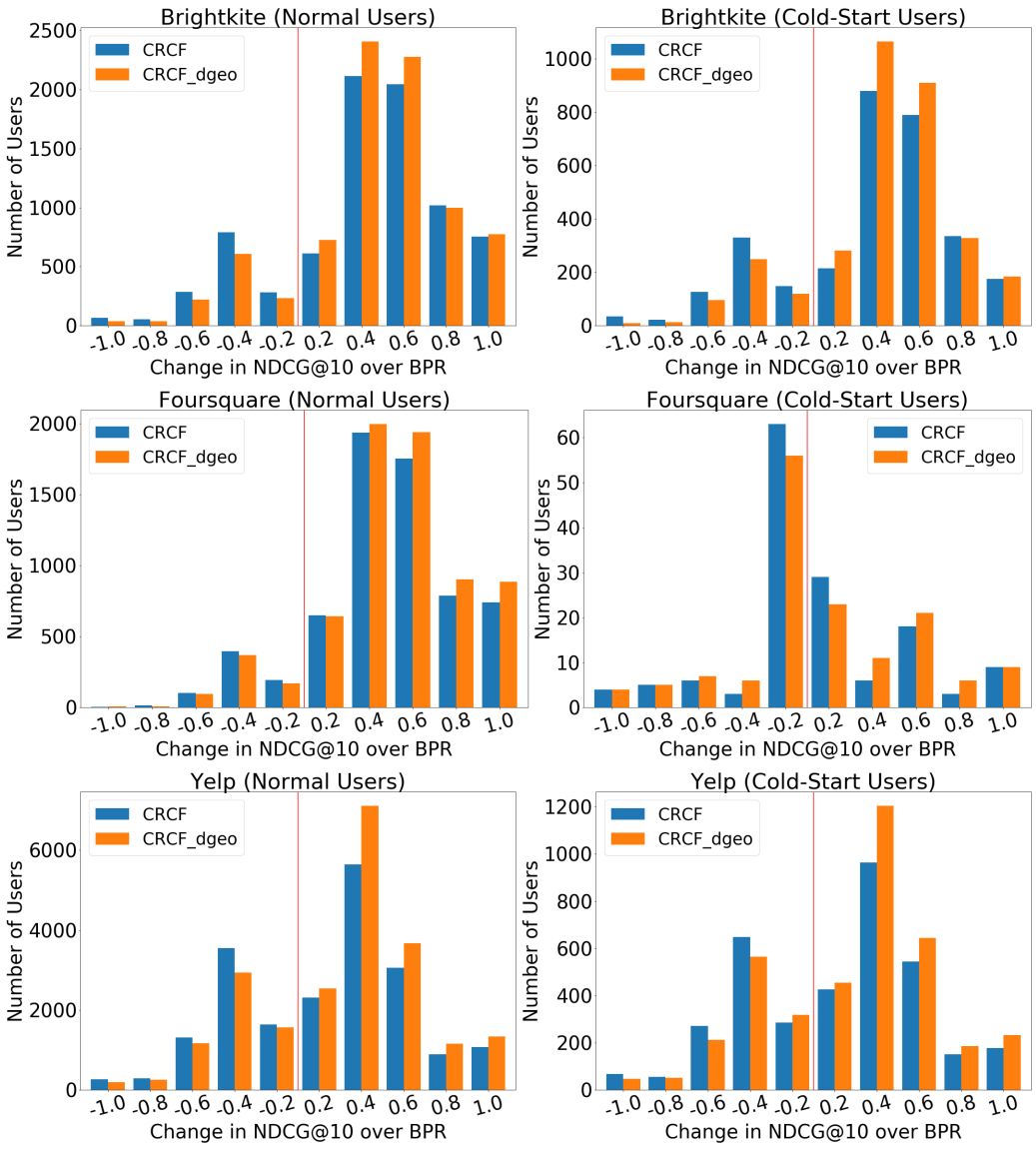


Figure 8: As per Figure 6, typical Wins-Losses histograms of the CRCF framework with or without the dynamic geo-based negative sampling approach (CRCF$_{dgeo}$ and CRCF) in comparison with the CF ranking baseline model, BPR, under the *Normal* and *Cold-Start Users* experiments.

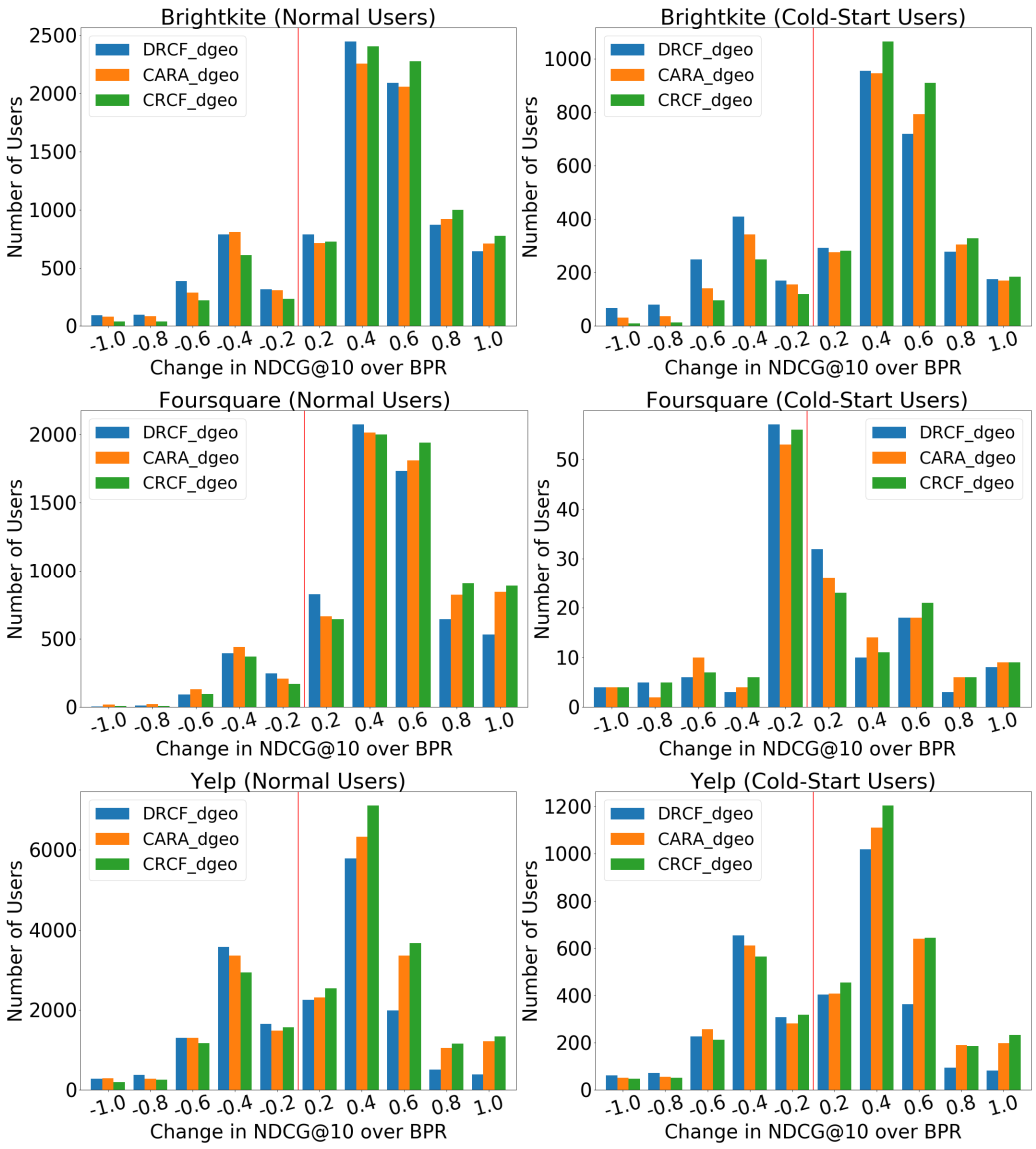


Figure 9: As per Figure 6, typical Wins-Losses histograms of target new models that incorporate the dynamic geo-based negative sampling approach (DRCF_{dgeo}, CARA_{dgeo} and CRCF_{dgeo}) in comparison with the CF ranking baseline model, BPR, under the *Normal* and *Cold-Start Users* experiments.

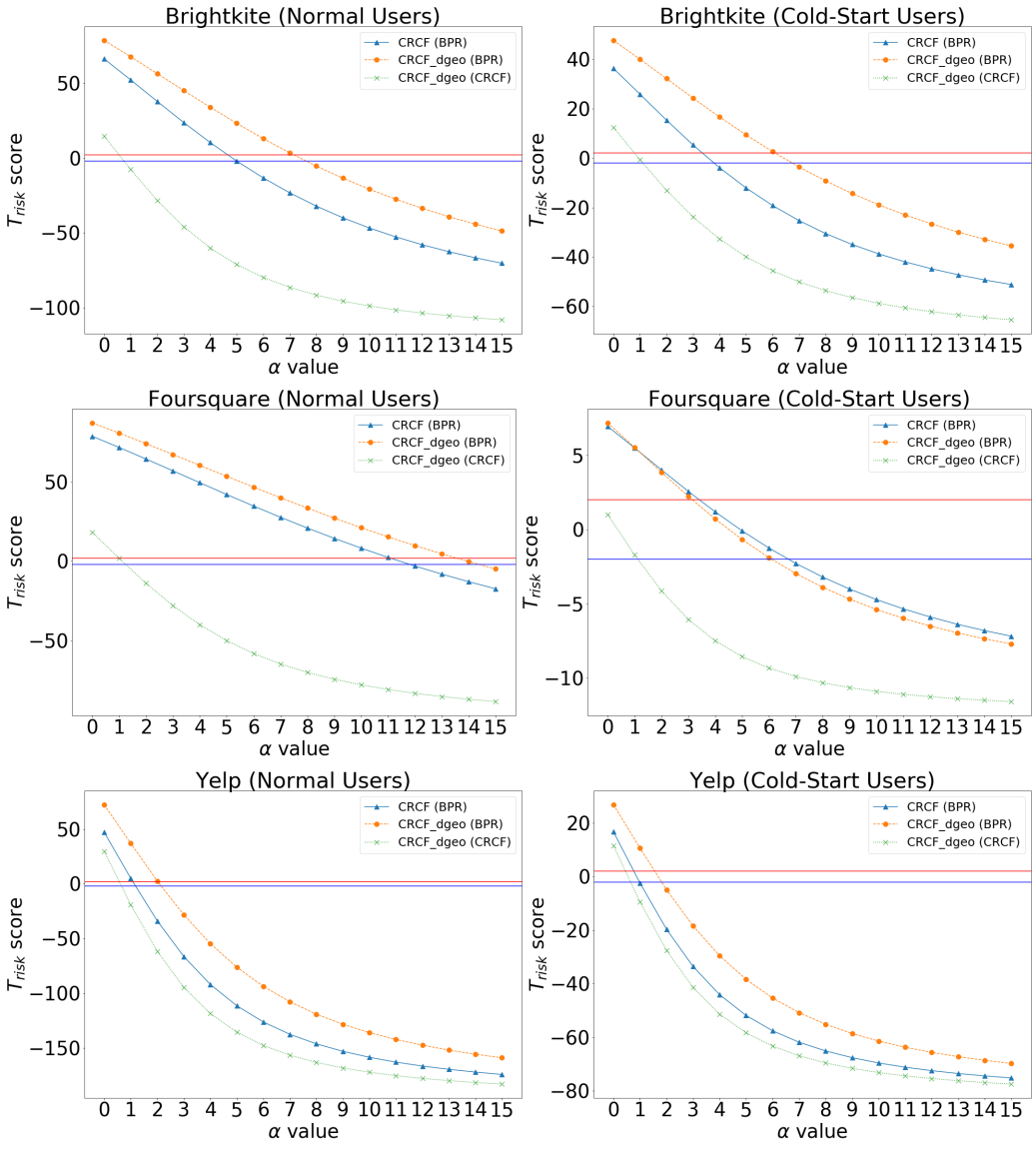


Figure 10: The change in standardised T_{risk} scores for CRCF and CRCF$_{dgeo}$ with respect to the BPR model, denoted inside the parentheses, over different α values under the *Normal* and *Cold-Start Users* experiments.

In particular, the following detailed findings can be observed from our thorough experiments:

- The contextual information associated with the sequences of the users' checkins (e.g. the time interval and distance between two successive checkins) is important in enhancing the quality of context-aware venue recommendation. In particular, in Figure 5, we demonstrated that our proposed CRCF framework, which exploits the state-of-the-art CARA architecture to leverage the contextual information, can outperform both the DRCF framework and the CARA architecture across the three used datasets. Moreover, the experimental results in Table 3 showed that the CRCF framework significantly outperforms the DRCF framework, which does not take the contextual information into account, across the three used datasets in terms of the HR and NDCG measures.

- We have demonstrated that the *dynamic* geo-based negative sampling approach, denoted with the $_{dgeo}$ suffix, can improve the effectiveness and robustness of various state-of-the-art context-aware recommendation approaches, and can alleviate the cold-start problem. In particular, CRCF$_{dgeo}$ exhibited a less significant risk of underperforming for a given user compared to BPR (Tables 7 & 8 and Figures 8 & 9).

- We have shown that leveraging the sequential order of users' checkins as well as the geographical information of venues can significantly improve both the effectiveness and robustness of the CRCF framework. In particular, CRCF$_{dgeo}$, the CRCF framework with the *dynamic* geo-based negative sampling approach, obtained over 6%, 3% and 11% improvements in terms of HR@10 over DRCF without the *dynamic* geo-based negative sampling approach (Table 4).

- In term of Wins, 45%, 54% and 33% of the *Normal* users in Brightkite, Foursquare and Yelp, respectively, received better venue suggestions from our proposed CRCF framework compared to the BPR model. In terms of Loss, only 10%, 6% and 18% of the users in Brightkite, Foursquare and Yelp, received less effective venue recommendations from CRCF compared to the BPR model (Table 5). In addition, the CRCF framework can generate more effective venue suggestions than

the BPR model for 42% and 32% of the *Cold-Start* users on Brightkite and Yelp, respectively (Table 6).

As future work, we plan to extend the CRCF framework to incorporate additional information such as the social relationships between users as well as the textual content of comments to further improve the quality of recommendation for CAVR - indeed, this has previously been shown to be useful for regularisation [38, 39].

References

[1] J. Manotumruksa, C. Macdonald, I. Ounis, A deep recurrent collaborative filtering framework for venue recommendation, in: Proc. of CIKM, 2017.

[2] J. Manotumruksa, C. Macdonald, I. Ounis, A contextual attention recurrent architecture for context-aware venue recommendation, in: Proc. of SIGIR, ACM, 2018, pp. 555–564.

[3] Y. Koren, R. Bell, C. Volinsky, Matrix factorization techniques for recommender systems, Computer (2009).

[4] F. Yu, Q. Liu, S. Wu, L. Wang, T. Tan, A dynamic recurrent model for next basket recommendation, in: Proc. of SIGIR, 2016.

[5] S. Tang, Z. Wu, K. Chen, Movie recommendation via blstm, in: Proc. of ICMM, 2017.

[6] Y. Zhang, H. Dai, C. Xu, J. Feng, T. Wang, J. Bian, B. Wang, T.-Y. Liu, Sequential click prediction for sponsored search with recurrent neural networks, in: Proc. of AAAI, 2014.

[7] S. Rendle, Factorization machines with libfm, ACM Transactions on Intelligent Systems and Technology (TIST) (2012).

[8] C. Cheng, H. Yang, M. R. Lyu, I. King, Where you like to go next: Successive point-of-interest recommendation., in: Proc. of IJCAI, 2013, pp. 2605–2611.

[9] H.-T. Cheng, L. Koc, J. Harmsen, T. Shaked, T. Chandra, H. Aradhye, G. Anderson, G. Corrado, W. Chai, M. Ispir, et al., Wide & deep learning for recommender systems, in: Proc. of DLRS, ACM, 2016, pp. 7–10.

[10] X. He, L. Liao, H. Zhang, L. Nie, X. Hu, T.-S. Chua, Neural collaborative filtering, in: Proc. of WWW, 2017.

[11] X. He, T.-S. Chua, Neural factorization machines for sparse predictive analytics, in: Proc. of SIGIR, 2017.

[12] Y. Zhu, H. Li, Y. Liao, B. Wang, Z. Guan, H. Liu, D. Cai, What to do next: Modeling user behaviors by time-lstm, in: Proc. of IJCAI, 2017.

[13] A. Beutel, P. Covington, S. Jain, C. Xu, J. Li, V. Gatto, H. Chi, Latent cross: Making use of context in recurrent recommender systems, in: Proc. of WSDM, 2018.

[14] E. Smirnova, F. Vasile, Contextual sequence modeling for recommendation with recurrent neural networks, in: Proc. of the DLRS, 2017.

[15] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural computation 9 (1997).

[16] J. Chung, C. Gulcehre, K. Cho, Y. Bengio, Empirical evaluation of gated recurrent neural networks on sequence modeling, in: Proc. of NIPS, 2014.

[17] D. Neil, M. Pfeiffer, S.-C. Liu, Phased lstm: Accelerating recurrent network training for long or event-based sequences, in: Proc. of NIPS, 2016.

[18] H. Jing, A. J. Smola, Neural survival recommender, in: Proc. of WSDM, 2017.

[19] Q. Liu, S. Wu, L. Wang, T. Tan, Predicting the next location: A recurrent model with spatial and temporal contexts, in: Proc. of AAAI, 2016.

[20] B. Twardowski, Modelling contextual information in session-aware recommender systems with neural networks, in: Proc. of RecSys, 2016.

[21] S. Rendle, C. Freudenthaler, Z. Gantner, L. Schmidt-Thieme, Bpr: Bayesian personalized ranking from implicit feedback, in: Proc. of UAI, 2009.

[22] F. Yuan, G. Guo, J. Jose, L. Chen, H. Yu, Joint geo-spatial preference and pairwise ranking for point-of-interest recommendation, in: Proc. of ICTAI, 2016.

[23] J. Manotumruksa, C. Macdonald, I. Ounis, A personalised ranking framework with multiple sampling criteria for venue recommendation, in: Proc. of CIKM, 2017.

[24] L. Wang, P. N. Bennett, K. Collins-Thompson, Robust ranking models via risk-sensitive optimization, in: Proc. of SIGIR, ACM, 2012, pp. 761–770.

[25] B. T. Dinçer, C. Macdonald, I. Ounis, Hypothesis testing for the risk-sensitive evaluation of retrieval systems, in: Proc. of SIGIR, ACM, 2014, pp. 23–32.

[26] L. Yao, Q. Z. Sheng, Y. Qin, X. Wang, A. Shemshadi, Q. He, Context-aware point-of-interest recommendation using tensor factorization with social regularization, in: Proc. of SIGIR, 2015.

[27] R. Deveaud, M.-D. Albakour, C. Macdonald, I. Ounis, On the importance of venue-dependent features for learning to rank contextual suggestions, in: Proc. of CIKM, 2014.

[28] R. Deveaud, M.-D. Albakour, C. Macdonald, I. Ounis, Experiments with a venue-centric model for personalised and time-aware venue suggestion, in: Proc. of CIKM, 2015.

[29] J.-D. Zhang, C.-Y. Chow, Geosoca: Exploiting geographical, social and categorical correlations for point-of-interest recommendations, in: Proc. of SIGIR, 2015.

[30] S. Zhao, T. Zhao, H. Yang, M. R. Lyu, I. King, Stellar: Spatial-temporal latent ranking for successive point-of-interest recommendation, in: Proc. of AAAI, 2016.

[31] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proc. of CVPR, 2016.

[32] H. Zhang, Y. Yang, H. Luan, S. Yang, T.-S. Chua, Start from scratch: Towards automatically identifying, modeling, and naming visual attributes, in: Proc. of MM, 2014.

[33] Y. Kim, Convolutional neural networks for sentence classification, arXiv preprint arXiv:1408.5882 (2014).

[34] Q. Liu, S. Wu, D. Wang, Z. Li, L. Wang, Context-aware sequential recommendation, in: Proc. of ICDM, 2016.

[35] Y. K. Tan, X. Xu, Y. Liu, Improved recurrent neural networks for session-based recommendations, in: Proc. of DLRS, 2016.

[36] D. Kingma, J. Ba, Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980 (2014).

[37] X. He, H. Zhang, M.-Y. Kan, T.-S. Chua, Fast matrix factorization for online recommendation with implicit feedback, in: Proc. of SIGIR, 2016.

[38] S. Liu, I. Ounis, C. Macdonald, Social regularisation in a BPR-based venue recommendation system, in: Proc. of FDIA, 2019.

[39] J. Manotumruksa, C. Macdonald, I. Ounis, Regularising factorised models for venue recommendation using friends and their comments, in: Proc. of CIKM, 2016.