2015 IEEE International Conference on Big Data (Big Data)

Multi-modal Learning for Video Recommendation based on Mobile Application Usage

Xiaowei Jia, Aosen Wang, Xiaoyi Li, Guangxu Xun, Wenyao Xu, and Aidong Zhang
School of Computer Science and Engineering
State University of New York at Buffalo, Buffalo, NY, USA, 14260-1660
Email: {xiaoweij,aosenwan,xiaoyili,guangxux,wenyaoxu,azhang}@buffalo.edu

***Abstract*—The increasing popularity of mobile devices has brought severe challenges to device usability and big data analysis. In this paper we investigate the intellectual recommender system on cell phones by incorporating mobile data analysis. Nowadays with the development of smart phones, more and more applications have emerged on various areas, such as entertainment, education and health care. While these applications have brought great convenience to people's daily life, they also provide tremendous opportunities for analyzing users' interests. In this work we develop an Android background service to collect the user behaviors and analyze their preferences based on their Android application usage. As one of the most intuitive media for visual representation, videos with various types of contents are recommended to users based on a proposed graphical model. The proposed model jointly utilizes the textual descriptions of Android applications and videos, as well as the extracted video content based features. Besides, by analyzing the user's habit of application usage we seamlessly integrate the user's personal interests during the recommendation. The extensive comparisons to multiple baselines reveal the superiority of the proposed model on the recommendation quality. Furthermore, we conduct experiments on personalized recommendation to demonstrate the capacity of the proposed model in effectively analyzing the user's personal interests.**

***Keywords*-mobile data; video recommendation; personalized recommendation;**

I. Introduction

The surge in mobile devices has provided tremendous opportunities for mobile data analysis [1]–[4]. While the recent development in hardware and system leads to the substantial improvement on cell phone performance, the progress in intellectual management acts as a catalyst in furthering technological advancements in mobile data analysis and personalized recommendation [1], [5]–[9]. On one hand, the intellectual management properly selects and recommends contents according to users' preferences, and consequently greatly improves the user experience on mobile devices. On the other hand, the simplified operations assist in reducing the fatigue of users nowadays from the large amount of diversified resources and contents on mobile phones.

In this paper we investigate the multi-modal learning on mobile data, which is a key consideration pertaining to the intellectual management. Moreover, we implement, to our best knowledge, the first Android recommender system that purely relies on users' habit of application usage. Given the popularity of smart phones, it poses a challenge how to effectively analyze the large amount of usage data to conduct the recommendation. On the other hand, due to the privacy issue, in most cases the developers are only allowed to access a rough description of users' behaviors on applications. Therefore it requires that we can well leverage the available information to analyze the users' preferences. Although the rough description only provides limited information, we can enhance the learning process by collecting various types of data online, e.g., we can crawl the textual description of a specific Android application on Google Play. Meanwhile, with the information of application usage a well-designed intellectual recommender system should be capable of recommending various contents (e.g. news, movies, pics, etc.) to the user rather than just new applications. Therefore we need to carefully investigate the relationships between the data in different modalities (e.g. text, images, and videos).

Based on the aforementioned analysis we can consequently recommend diversified contents to users, as shown in Figure 1. While the proposed work in this paper can be easily adapted to different data formats, in our implementation we concentrate on the recommendation of video, which serves as one of the most intuitive media for visual representation. By developing an Android application (with a background service) to track and analyze the users' application usage, we can recommend online videos accordingly.

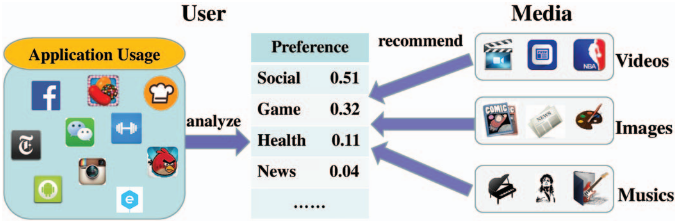


Figure 1: The recommendation of various media based on application usage.

In our developed Android application, the users are required to rate each recommended videos from two aspects - whether the video fits their habit of application usage (rating of relevance), and whether they like the video (rating of interests). While the rating of relevance indicates the

978-1-4799-9926-2/15/$31.00 ©2015 IEEE

matching relationship between the recommended video and the user's habit of application usage, the second rating aims at reflecting the user's real feeling on the video. The difference between two rating scores can be illustrated in a simple scenario: while the participant i is currently using some applications on English education, he is not interested in this area (e.g. his parents force him to use such applications). So given a recommended video on health care, he may give a high rating of relevance but a low rating of interests. During the recommendation, we should pay more attention on the rating of interests which can truly reflect the user's real feeling. However, as observed from our collected data, in most cases the rating of interests is still related with the rating of relevance. Given a recommended video, most users will get surprised and excited if the video very well matches his recent habit of application usage.

After gathering the feedback from the users, we propose a novel graphical model which utilizes the heterogeneous information from text and videos to jointly learn the relationship between the users' behaviors and their preferences on different videos. According to our study, the final rating of interests is mainly relevant with two factors: the rating of relevance and the user's long-term preferences. To this end we propose a two-stage training structure where we first consider the training on relevance and subsequently conduct the training on the user's interests.

II. Problem Definition

In this work, we develop an Android application with a background service to track the usage of mobile applications among a group of volunteers. With the collected information, we are able to analyze their preferences over various types of contents, such as games, news, and health care. Then we crawl a set of 2-3 minutes videos from websites and recommend to the user. For each recommended video the user will be asked to rate with respect to two questions: (1) whether he likes the video (regarding interests), and (2) whether he feels the recommended video is related with his habit of using mobile applications (regarding relevance). The answer for each question falls in five degrees (e.g. in the first question, 5 means "strongly like", 1 means "totally dislike"). More formally, for each user i we track all his used applications which can be represented as a series $\{A_1^i, A_2^i, ..., A_{K^i}^i\}$. We also record the time that the user spends on each application $\{t_1^i, ..., t_{Ki}^i\}$. It is worthwhile to note that users may repeat using same applications, and K^i stands for the total times that the user i launches an application. Besides the application records, we collect the answers from the user i on the two given questions, $\{I_1^i, ..., I_{V^i}^i\}$ (regarding interests) and $\{R_1^i, ..., R_{V^i}^i\}$ (regarding relevance), where V^i represents the total number of videos watched by the user i. During the training procedure, we wish to learn the relationship between the usage of the applications and the preference on different videos for each user. After obtaining the learned model, we expect to predict the degree of both relevance and interests based on the user's habit of application usage, and consequently conduct the high-quality recommendation.

III. Methodology

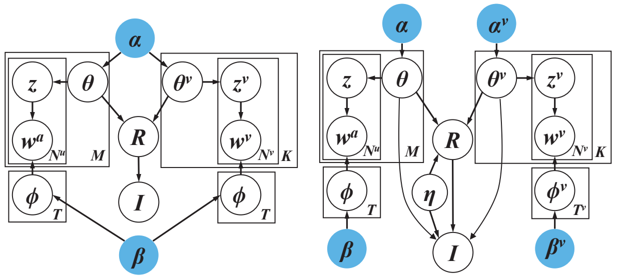
The applications serve as the most attractive and convenient utilities on mobile phones, and cover a variety of areas, including social services, gaming, health care, e-commerce, etc. According to our study, we find that most users spend lots of time on using mobile applications and each user usually pays more attention to several specific types of applications. Therefore we propose to explore the user's preferences by tracking their usage of the applications.

A. Multi-modal Generative Modeling

To analyze the users' preferences through their application usage records, we propose to utilize the textual description of each application from Google Play. Then we set a fixed time interval and concatenate the descriptive text of the applications involved in consecutive intervals as the user's "document". Hence the more time the user spends on a specific application, the larger weight this application will be given in the training. Likewise, to categorize each video, we treat the descriptive text for each video as a "document". For most online videos, the textual information can be obtained from the web page, including the description provided by the uploader and the comments.

As we can access the description for both applications and videos, the most intuitive solution is to infer the corresponding relationships between the user's document and the video's document. As shown in Figure 2 (a), this naive model considers the textual description of applications and videos in a unified framework. For each word in user's document, we first sample the topic distribution θ from the Dirichlet distribution $Dir(\alpha)$, and then select a topic z from the multinomial distribution $Mult(\theta)$. Meanwhile, the word distribution ϕ is sampled for each topic, and a word is selected from the word distribution of the selected topic. On the other hand, for each word in video description we sample the video's topic distribution θ^v from the same Dirichlet distribution $Dir(\alpha)$, and select a topic z^v. Then the word is generated in the similar fashion with the user's document. Finally, the rating of relevance R is generated based on the logistic function of the inner product of θ and θ^v, as $logistic(a(\theta\theta^v)+b)$, and subsequently the rating of interests is sampled from a Gaussian distribution $I \sim \mathcal{N}(R, \gamma^2)$.

However, the above naive model fails to include many key factors that are relevant to users' feelings and interests, which leaves our problem far from well solved. To this end we propose a revised graphical model, as shown in Figure 2 (b). First, as it is noticeable that the textual descriptions of Android applications and videos usually cover different aspects, we separately extract topics from the documents



(a) (b)

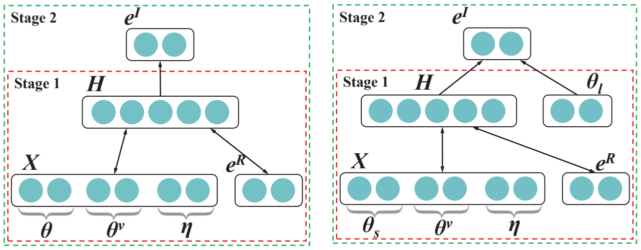
Figure 2: The graphical models for recommendation. (a) The naive text matching model. (b) The proposed graphical model. The variables in the lower right corner of each box refer to the number of samples for the sampling.

of users and videos. For each word in user's document and video description, we sample the topic distribution θ and θ^v respectively from two different Dirichlet distributions $Dir(\alpha)$ and $Dir(\alpha^v)$. Similarly the word distributions for user's document and video's document are generated following two different Dirichlet distributions $Dir(\beta)$ and $Dir(\beta^v)$, and subsequently the word is sampled based on the selected topic. The topic extraction can be conducted following a Gibbs sampling procedure [10].

Besides the descriptive text, in the proposed model we integrate the video content based features η, as the descriptive text of a video sometimes cannot fully reflect its content. These features η include the average contrast ratio, color histogram, averaged frame image, the variance across frames (to measure the smoothness), detected scenes and the video properties such as video duration, frame rate, average mixed bit rate, etc. By combining the user's topic distribution θ, the video's topic distribution θ^v and the video features η, we generate the rating of relevance R. Finally, since most users have preference on specific topics, we propose to jointly utilize θ, θ^v, η, and R, to generate the rating of interests I.

B. Relevance and Interests

To solve the classification of relevance and interests, we propose a two-layer learning model, as shown in Figure 3 (a). The training is conducted in two stages. In stage 1, the first layer is trained as a generative Restricted Boltzmann Machine (RBM) [11] which aims as classifying the relevance R. Then in stage 2, the learned parameters from the first stage are used to initialize the two-layer neural network and the supervised training is conducted for the classification of interests I. From Figure 3 (a) we can observe that the interests I depend on both the personalized features X (including user topic distribution θ, video topic distribution θ^v and video content based features η) and the relevance rating. The idea can be well illustrated in a real scenario: a user usually judges a video from two aspects, whether the recommended video is related to his recent operations on the mobile device (i.e. relevance), and whether he is interested in the content (i.e. personalized features X). In this section we will introduce the two-stage training of the model. While it is discussed as a two-layer neural network, in practice it can be easily extended to a deep structure for higher accuracy.



(a) (b)

Figure 3: The two-layer learning model. (a) the learning model of relevance and interests. (b) the model with the long-term personalized interests.

In the generative RBM, the visible layer consists of both the features X and the label units (i.e. the one-hot representation of relevance rating $\mathbf{e}^R$). The hidden layer is used to capture the relationship between the features and the label units. The loss function is defined as:

$$L_{gen} = -\sum_{Data} logP(X,R) = -\sum_{Data} log \sum_{H} \frac{exp(-E(R,X,H))}{Z}, \quad (1)$$

where Z represents the normalization factor, and the energy function of the generative RBM is defined as:

$$E(R,X,H) = -H^TWX - H^TU\mathbf{e}_R - B^TX - C^TH - D^T\mathbf{e}_R, \quad (2)$$

where W represents the weights between hidden units H and feature units X, U represents the weights between H and label units, B, C and D serve as biases for X, H and label units respectively, and $\mathbf{e}_R$ stands for the one-hot vector with the R^{th} position set as 1. Compared with traditional RBM [12], the generative RBM model involves the connection between H and label units by a weight matrix U, and the bias of label units D. Besides, the optimization procedure with respect the joint distribution $P(X,R)$ has proven to be able to greatly reduce the overfitting problem [11].

In the generative RBM model, the gradient of parameters $\Theta = \{W,U,B,C,D\}$ can be computed as:

$$\begin{aligned}\frac{\partial logP(X,R)}{\partial\Theta} &= -\mathbb{E}_{H|X,R}[\frac{\partial}{\partial\Theta}E(R,X,H)] \\ &+ \mathbb{E}_{R,X,H}[\frac{\partial}{\partial\Theta}E(R,X,H)],\end{aligned} \quad (3)$$

where the second term on the right side represents the model expectation. Due to its computational intractability, we solve it using Contrastive Divergence algorithm [12], which is in effect a stochastic approximation of the gradient by following the alternative Gibbs sampling.

After the training of the generative RBM, the learned parameters are used to initialize the weights of the first layer in the neural network. Then the two-layer network is trained to classify the rating of interests. The input consists of the input features X and the relevance ratings R, and the output is the one-hot representation of the interests e^I. The prediction of interests can be conducted by following:

$$P(I_i = 1|H) = sigm(W_{i.}^2 H + D_i^2), \tag{4}$$

where W^2 and D^2 represent the weights and bias for the second layer.

C. Short-term Habit and Long-term Preference

As we have discussed earlier, the rating of interests is determined by both the relevance and the personal preference. According to our collected data, we can observe that some participants may use a specific type of Android applications for a short period, while these applications do not match exactly their long-term interests. For instance, user i likes playing mobile games, and thus spends most time on gaming. However, during a specific time period he may frequently use some other types of applications, as his parents force him to learn some skills through the educational softwares or his friends introduce a new health care application to him. Although the user's preference on the recommended videos heavily relies on the correspondence between his recent short-term application usage and the video content, the user's long-term application usage can reflect his real interests which also impact how much he likes the video. To this end, we revise the proposed two-layer learning model by differentiating the user's short-term habit and long-term preference, as shown in Figure 3 (b). In this revised model the rating of interests depends on the short-term application usage, the video features, the rating of relevance and the long-term preference. In our implementation, the long-term usage covers the applications used in the past week while the short-term usage only covers the past 5 hours. To keep the the user's interests updated, we retrain the long-term preference every week. The intuitive idea behind this can be illustrated in a real scenario: user i is a fan of mobile games. However, recently he starts to frequently use the applications on social interactions, such as Facebook. Gradually he finds the social activities more interesting and spends less time on the gaming. Hence by regularly updating the long-term information we can detect the case where the user alters his long-term preference.

IV. Experiments

A. Dataset Description

Synthetic Dataset: We validate the proposed method on a synthetic dataset with 1000 users. To simulate the real scenarios, we construct user and video clusters where the objects in each cluster share similar attributes. Then we define a vocabulary and generate the description based on multinomial sampling.

Real-world Dataset: We dispatch test phones and record the volunteers' usage of applications. Then we crawl a set of 2-3 minutes videos from `Youtube.com` and recommend to users. During the application recording and video recommendation, we mainly focus our attention on 11 aspects: health and fitness, entertainment, shopping, cooking, news, music, photo shooting, social behaviors, English education, gaming and racing games. For each recommended video, the user is required to provide the ratings of relevance and interests.

Table I: Prediction of relevance (R) and interests (I) on synthetic dataset using AUC score.

Methods	Test 1(R)	Test 2(R)	Test 1(I)	Test 2(I)
MF	0.701	-	0.683	-
MC	0.755	-	0.742	-
HR	0.703	0.684	0.699	0.671
HCM	0.729	0.711	0.705	0.685
TM	0.734	0.729	0.706	0.689
MVR	**0.766**	**0.751**	**0.760**	**0.733**

Table II: Prediction of relevance (R) and interests (I) on real-world dataset using AUC score.

Methods	Test 1(R)	Test 2(R)	Test 1(I)	Test 2(I)
MF	0.653	-	0.621	-
MC	0.686	-	0.675	-
HR	0.659	0.660	0.624	0.621
HCM	0.677	0.667	0.645	0.630
TM	0.673	0.667	0.611	0.579
MVR	**0.727**	**0.713**	**0.705**	**0.668**

B. Video Recommendation

We name our proposed model as **M**ulti-modal learning for **V**ideo **R**ecommendation (MVR), and we compare with five baseline methods, including Matrix Factorization (MF) [13], [14], Matrix Co-factorization (MC) [15], History based Regression (HR) [16], Hybrid Collaborative Model (HCM) [16] and Text Matching (TM) (depicted in Figure 2).

We separately conduct experiments on two test sets (Test 1 and Test 2). The users involved in the first test set have available video rating records in the training set, while the second test set consists of users without rating records in the training set. The results in AUC scores are shown in Table I and Table II. Some collaborative filtering based models such as MF and MC cannot be adapted to the second test set.

From Table I and Table II we can observe that the proposed method outperforms the baselines by a considerable margin. Compared with TM and MVR, we can conclude that it is necessary to separately model the textual description of videos and applications. Also, the video features are crucial in cases where the textual description cannot fully cover the video content. On the other hand, the proposed model generates the rating of interests by combining the rating of relevance and the extracted features from user and videos,

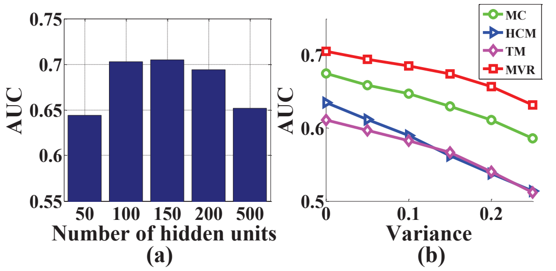


Figure 4: The performance on (a) parameter sensitivity test and (b) noise resistance test.

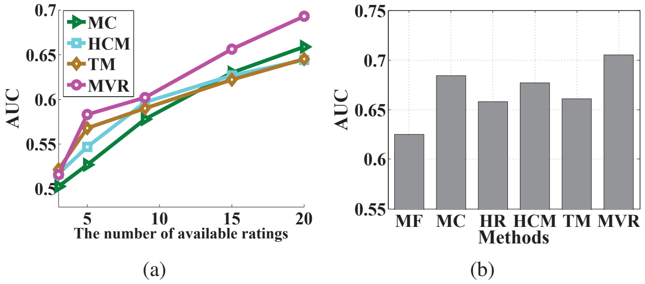


Figure 5: (a)The performance of interests prediction on cold start test. (b)The performance of personalized recommendation.

while the traditional models only utilize the features, and TM only utilizes the relevance ratings. Moreover, from the results on the two test sets, it is noticeable that the performance on the interests prediction drops more sharply from test set 1 to test set 2 compared to the decrease of relevance, which lies in that the users' interests I rely on more personalized factors. The models may fail to capture these factors without enough rating records from the same set of users during the training procedure.

Besides, we testify the proposed model in the cold start phase, which is a key measure for most recommender systems. In this test we select 30% users and restrict our available information on these users to only first M video ratings and the application usage records before they rate the M^{th} video. The value of M ranges from 5 to 20, and the prediction results under different values of M are shown in Figure 5 (a). The reduction of available ratings will directly impact the performance of the collaborative filtering based algorithms, which is verified in the figure that MC performs poorly at the beginning. As for the content-based methods, the reduction of available application usage records severely affects the quality of extracted user features. From the figure we can observe that our proposed model outpaces the baselines in the cold start test and shows faster recovery speed when gaining more training data. In practice, the proposed two-layer learning model can extract representative features during the training process, therefore greatly mitigates the overfitting given small samples.

In addition, we conduct the parameter sensitivity test by measuring the performance of interests prediction on test set 1 with different number of hidden units, as shown in Figure 4 (a). While the extremely small number of hidden units do not have enough capacity to reflect the latent relationships, the extremely large number of hidden units can easily result in overfitting. In Figure 4 (b), we demonstrate the robustness of our model by adding Guassian noise to the training data. Compared with baseline models, the performance of MVR shows a relatively gradual decline with the increasing noise.

C. *Personalized Recommendation*

To adapt the baseline models to personalized scenarios we wish to separately train the baseline models for different users. Unfortunately, as the collected dataset involves large amount of participants, it is very space-consuming and time-consuming to train separate models for each user. Therefore we first group the users into 25 clusters based on their usage of Android applications and then train different models for each user cluster. On the contrary, as our proposed model MVR has already taken account of the personalized information, we directly take all the collected data as input for the training.

Figure 5 (b) shows the comparison between the proposed model and the baselines on personalized recommendation. While our proposed model is directly trained on the whole dataset, we can observe that the performance is still better than the baselines which are trained separately on each user cluster. Hence we can conclude that the extracted long-term topic distribution can truly reflect the user's preferences. Compared with the original training process on the whole dataset, it can be noticed that the separate training only brings limited improvement or even a slight decrease to the collaborative filtering based methods. While the separate training on each user cluster facilitates the learning of personalized factors, it reduces the number of users involved in each collaborative filtering process. In contrast, the content-based methods demonstrate a substantial improvement with the separate training.

The extracted long-term interests play an important role in the personalized recommendation. In this part we measure the performance on real-world dataset by adjusting the time period covered by the long-term application usage, from the past 5 hours to all the available records. At the same time, we change the length of the short-term period, from 5 minutes to 1 week. Table III depicts the results in AUC score with different length of short-term period and long-term period. The values outside parenthesis denote the AUC of interests prediction while the values inside parenthesis denote the AUC of relevance. In the test we ignore the trivial cases where the short-term period is longer than long-term period. It is noticeable that the AUC for the relevance prediction stays the same for the entries in the same row as the relevance only relies on the short-term application usage. From the table, we can observe that the AUC score increases when the long-term period varies from 5 hours to 1

Table III: Prediction of interests (relevance) with different length of short-term and long-term usage. The horizonal line denotes the time period covered by the long-term application usage, while the vertical column denotes the time period covered by the short-term application usage.

AUC	5 hrs	12 hrs	2 days	1 week	2 weeks	all
5 min	0.612(0.702)	0.625(0.707)	0.657(0.707)	0.688(0.707)	0.691(0.707)	0.691(0.707)
30 min	0.609(0.702)	0.625(0.702)	0.655(0.702)	0.685(0.702)	0.688(0.702)	0.687(0.702)
2 hrs	0.617(0.719)	0.634(0.719)	0.658(0.719)	0.697(0.719)	0.697(0.719)	0.694(0.719)
5 hrs	0.621(0.727)	0.638(0.727)	0.667(0.727)	0.705(0.727)	0.702(0.727)	0.705(0.727)
12 hrs	-	0.631(0.702)	0.661(0.702)	0.686(0.702)	0.687(0.702)	0.690(0.702)
2 days	-	-	0.649(0.686)	0.671(0.686)	0.675(0.686)	0.679(0.686)
1 week	-	-	-	0.659(0.667)	0.661(0.667)	0.659(0.667)

week. This mainly stems from the fact that the longer period can better reflect the users' real interests. However, when it becomes longer than 1 week, the AUC reaches a plateau, or even shows a slight decrease. Such phenomenon lies in that the we can hardly detect the transition where the user alters his interests within an extremely long period. On the other hand, the short-term period that is longer than 5 hours results in a decrease of AUC in relevance prediction, and consequently a drop in interests prediction, since it cannot well capture the users' current habit of application usage. It is noteworthy that the selection of 5-min short-term period performs slightly better than that of 30-min period, which unveils that some users may focus on their most recent activities when they consider the relevance.

V. Conclusion

In this paper we proposed and implemented an innovative model for video recommendation on cell phones. The model effectively uncovered the users' preferences from their habit of application usage, and jointly utilized the knowledge from text and videos. Meanwhile, in the model we took account of both the long-term preference and the short-term habit in order to better simulate the users' judgement on relevance and interests. The extensive experiments showed the superiority of our model on effectively analyzing the user behaviors and making proper recommendation. In addition, the experimental results well supported our intuition in differentiating the short-term habit and long-term preference.

References

[1] V. W. Zheng, Y. Zheng, X. Xie, and Q. Yang, "Collaborative location and activity recommendations with gps history data," in *WWW*, 2010.

[2] N. Eagle and A. Pentland, "Reality mining: sensing complex social systems," *Personal and ubiquitous computing*, 2006.

[3] R. Montoliu and D. Gatica-Perez, "Discovering human places of interest from multimodal mobile phone data," in *Proceedings of the 9th MUM*, 2010.

[4] Y. Zheng, L. Zhang, X. Xie, and W.-Y. Ma, "Mining interesting locations and travel sequences from gps trajectories," in *WWW*, 2009.

[5] M.-H. Park, J.-H. Hong, and S.-B. Cho, "Location-based recommendation system using bayesian users preference model in mobile devices," in *UIC*, 2007.

[6] A. Karatzoglou, L. Baltrunas, K. Church, and M. Böhmer, "Climbing the app wall: enabling mobile app discovery through context-aware recommendations," in *Proceedings of the 21st ACM CIKM*, 2012.

[7] F. Ricci, "Mobile recommender systems," *ITT*, 2010.

[8] D. Quercia, N. Lathia, F. Calabrese, G. Di Lorenzo, and J. Crowcroft, "Recommending social events from mobile phone location data," in *Data Mining (ICDM)*. IEEE, 2010.

[9] E. Kaasinen, "User needs for location-aware mobile services," *Personal and ubiquitous computing*, 2003.

[10] M. Steyvers and T. Griffiths, "Probabilistic topic models," *Handbook of latent semantic analysis*, 2007.

[11] H. Larochelle, M. Mandel, R. Pascanu, and Y. Bengio, "Learning algorithms for the classification restricted boltzmann machine," *JMLR*, 2012.

[12] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural computation*, 2006.

[13] Y. Koren, R. Bell, and C. Volinsky, "Matrix factorization techniques for recommender systems," *Computer*, 2009.

[14] X. Jia, N. Du, J. Gao, and A. Zhang, "Analysis on community variational trend in dynamic networks," in *Proceedings of the 23rd ACM CIKM*, 2014.

[15] L. Hong, A. S. Doumith, and B. D. Davison, "Co-factorization machines: modeling user interests and predicting individual decisions in twitter," in *Proceedings of the sixth WSDM*, 2013.

[16] Y. Zhang, H. Wu, V. Sorathia, and V. K. Prasanna, "Event recommendation in social networks with linked data enablement." in *ICEIS (2)*, 2013.