Predicting User's Multi-Interests With Network Embedding in Health-Related Topics

Zhipeng Jin[1], Ruoran Liu[1], Qiudan Li[1] , Daniel D. Zeng[1,2], YongCheng Zhan[2], Lei Wang[1]
[1]The State Key Laboratory of Management and Control for Complex Systems
Institute of Automation, Chinese Academy of Sciences Beijing 100190, China
[2] Department of Management Information Systems, University of Arizona, Tucson, Arizona, USA
{jinzhipeng2013, liuruoran2016, qiudan.li, dajun.zeng, l.wang}@ia.ac.cn, yongchengzhan@email.arizona.edu

Abstract—With the rapid growth of Web 2.0, social media has become a prevalent information sharing and seeking channel for health surveillance, in which users form interactive networks by posting and replying messages, providing and rating reviews, attending multiple discussion boards on health-related topics. Users' behaviors in these interactive networks reflect users' multiple interests. To provide better information service for users, it is necessary to analyze the user interactions and predict users' multi-interests. Most existing work in predicting users' multi-interests based on multi label network classification focuses on using approximate inference methods to leverage the dependency information to improve classification results. Inspired by deep learning techniques, DEEPWALK learns label independent latent representations of vertices in a network using local information obtained from truncated random walks, which provides an efficient way for predicting users multi-interests from user interactions. In this paper, we develop a user's multi-interests prediction model based on DEEPWALK, weight information of user interactions is considered when modeling a stream of short constrained random walks and SkipGram is employed to generate more accurate representations of user vertices, which help identify users' interests. Experimental results on two real world health-related datasets show the efficacy of the proposed model.

Keywords—user interaction network; multi-interests prediction; weight information; DEEPWALK

I. INTRODUCTION

The rapid growth of Web 2.0 has made social media a significant platform for health surveillance[1] and social intelligence[2], in which users form interactive networks by posting and replying messages, providing and rating reviews, attending multiple discussion boards. Take Reddit, one of the most popular forums in the world, as an example, it has recently become an important data source for analyzing health related topics. Pavalanathan and Choudhury used Reddit to study mental health [3]. Arthur used Reddit to track the 2014 Ebola outbreak [4]. Wang[5] showed that Reddit is heavily used by the e-cigarette and vaping community to share information about flavors and other aspects of e-cigarette use, thus can be mined for valuable information on self-reported e-cigarette flavor use. On Reddit, users are allowed to create communities (called "subreddits") where they can perform various social actions reflecting their interests, such as posting, replying, voting on their interested messages. Based on social network theory[6, 7], these actions performed can build social links among users and can be recorded using an interactive graph. A link between user a and b is regarded as these two users performing the same actions on the same message, thus holding similar interests. The number of same messages these two users have replied or voted may determine how similar they are to each other. A user may be engaged in several communities simultaneously, the multiple communities a user has joined reflect the users' multi-interests, which can be used as users' interest labels. To provide better information service for users, it is necessary to analyze the user interactions graph and predict users' multi-interests.

Decision-making departments may track the user interaction network and the multiple interests for an informed policy, user behaviors and corresponding interested labels could act as a feedback channel to an announced policy, thus the departments could track the messages and adjust their policies accordingly. An individual user may track other people's interests and gain information about it.

However, most existing work in predicting users' multi-interests based on multi label network classification focuses on using approximate inference to improve classification results. Recently, inspired by deep learning techniques, a novel approach named DEEPWALK has been proposed to learn label independent latent representations of vertices in a network using local information obtained from truncated random walks, the method provides an efficient way for predicting users multi-interests from user interactions.

In this paper, we focus on learning users' latent representations from interactive networks and predicting users' multiple interests based on DEEPWALK. To reflect dynamic communication behavior of users and capture the interactive effect between interaction network structure and users' actions, a novel approach taking into consideration the weight information of the link is proposed to generate more accurate representation of user vertices, which helps identify users' interests.

We empirically evaluate the performance of the proposed model on two real world health-related datasets. Experimental results show that taking information on edge weights into consideration could allow us make more accurate analysis of users' multi interests. The mined user interests could serve as a

978-1-5090-0620-5/16/$31.00

product feedback channel for both businesses and consumers. A consumer could rely on the interests when making an informed decision whether to buy a product. The firms could better understand users' interests on their products, and take specific measures to improve their services.

The rest of this paper is organized as follows: In section II, we discuss relevant studies in the literature. The detailed procedure of our model is presented in Section III. We empirically evaluate our algorithm in Section IV. Section V sums up our study and discusses future research directions.

II. Literature Review

Our work is related to health surveillance using social media, analysis and prediction of user interest, deep learning and DeepWalk. In this section, we review the related works.

A Health surveillance Using Social Media

Social media such as Facebook, Twitter, and YouTube have recently become a significant platform for health surveillance and social intelligence [1, 2]. With the prevalence of e-cigarettes, discussions of e-cigarette benefits, risks, and effects on health have become a hot topic [5, 8, 9]. In general, research on Twitter focused on detection of trends and patterns. YouTube, on the other hand, provides rich information in videos. Facebook is good at social network construction.

E-liquid, or e-juice, is regarded as the main component in e-cigarettes, which is a mixture of propylene glycol (PG), vegetable glycerine (VG), flavor extracts, and nicotine [10]. The flavors are usually natural or artificial flavor concentrates generally recognized as safe. They tend to fall into a few categories: Fruits, Beverages, Sweet, Nuts, Cream, Menthol, Seasonings, and Tobacco[5]. Tobacco companies have successfully marketed traditional tobacco products to youth by using flavor varieties. For instance, Kostygina et al.[11] find that menthol and candy-like flavors increased little cigars' and cigarillos' appeal to starters by masking the heavy cigar taste, which might be a potential factor to result youth use. Although the FDA banned flavored tobacco in 2009 [12], flavors are still widely used for e-cigarettes. Etter et al.[13] find that Menthol flavor can relieve the craving to smoke, possibly by obtaining a better throat hit. These studies showed that e-cigarette flavors could be dangerous but attractive, which should be carefully studied by researchers.

Reddit[1] and JuiceDB[2] provide very interesting discussion and review services for e-cigarette juice, where users can review their interested e-liquid, join their interested community, and reply their interested messages, etc. All of the above behaviors reflect people's vaping experience and interest in e-juice. For example, on Reddit, users who are interested in policy of banning juice flavor may join in community about "Policy", users that are fond of mixing different flavors may join in "DIY juice" community, etc. Predicting which community a user will join in will obtain users' preference on juice. On JuiceDB, each review is accompanied by an overall rating, detailed flavor descriptions of a juice, by analyzing the user's review interaction network, we can gain better understanding of users' preference on juice flavor. The analysis of the users' interest in e-liquid could open a window for us to understand the behavior of e-cigarette users, thus the promotion strategies and regulation could be developed more pertinently.

However, little research in the field of social media has paid attention to predicting users' interest in flavor of e-liquid and user experience in the use of e-cigarette. This paper aims to gain a systematic understanding of users' preference of e-liquid flavor and user experience of using e-cigarette, by analyzing e-cigarette related interaction network on Reddit and JuiceDB.

B User Interest Analysis and Prediction

User interest analysis and prediction play important roles in helping understand user behaviors and providing better information services for users. Existing works on user interest analysis mainly focus on mining users' interests from user-generated contents or users' relationship with each other.

Han et al.[14] express user interest in terms of probabilities that a user has interest in categories. A TF-ICF method and the topic modeling method are combined to extract explicitly or implicitly presented features of categories, then features from news categories and user messages are compared to infer user interest. Lappas et al.[15] capture users' interests by modelling social endorsement network and extract relevant and descriptive tags for entities they endorse. Bhattacharya et al.[16] observe that a user usually follows experts on different topics of her interest to get the information on those topics and thus propose to discover users' interests from the topical expertise of the users they follow. They deduce the topical expertise based on a social annotations system Twitter Lists and predict the interested topics through the experts a user subscribes to.

As for relationship-based methods, in social media, there exist multiple types of social relationships between users. For example, on Twitter, following, retweeting and mentioning are three major types of social links between two users. As shown in Welch et al. 'research[17], these links may indicate different levels of topical relevance, retweeting is a stronger indicator of topical relevance than following. He et al.[18] utilize user's following relationship information and propose Bi-Labeled LDA to capture interest tags for non-famous user through their relationship with famous users in Twitter. Wang et al.[19] argue that learning user interests should be based on the link structure assumption, which can be potentially more robust to adapt to various types of social connections and more resilient to sparse and dynamic networks. Under this assumption, node similarities are measured based on the local link structures between two nodes. For example, people sharing many followers or followees are likely to be similar in terms of their topical interests. They propose a regularization-based framework by utilizing the relation bipartite graph, which consists of out-link regularization and in-link regularization. Bao et al.[20] propose a temporal and social probabilistic matrix factorization model to predict users' potential interests in micro-blogging, which provides a unified way to fuse the time information and the social network structure to predict user interest. Ma et al.[21] study the relationship between the user's trust network and the user-item matrix systematically

[1] https://www.reddit.com/

[2] https://www.juicedb.com/

and propose a method integrating social network structure and the user-item rating matrix. Jamali et al.[22] propose a SocialMF model by incorporating trust propagation into a matrix factorization for recommendation in social network.

C Deep Learning and DeepWalk

Deep-learning techniques[23] are representation-learning methods, which can automatically discover multiple levels of representations from raw data. They have proven successful in various fields such as computer vision[24], speech recognition[25], and natural language processing[26]. Perozzi et al.[27] firstly introduce deep learning into network analysis and propose DeepWalk algorithm to learn social representations of a graph's vertices, by modeling a stream of short random walks. It generalizes neural language models to process a special language composed of a set of randomly-generated walks.

Based on the above interesting research work, our work constructs user interaction network based on link structure assumption, then learns the user latent representation by combining weight information and DeepWalk. Finally, the proposed model is applied to predict users' interests in juice flavor and user experience in using e-cigarette. The mined social media data-driven scientific findings could have significant benefit to regulatory agencies like the FDA so that they can develop a better understanding of marketing and media that could come under their regulatory review.

III. Proposed Model for Users' Multi-Interests Prediction

A. Theoretical Background

Users' interest in social media can be well reflected by user-generated contents and users' interaction relationships. If a user is interested in some topics, she/he may publish, reply or vote on messages about those topics or follow other users who publish them[28]. It has been pointed out by social network researchers that the social structure of a social network is more indicative of certain actions of users than the attributes of individuals involved [6, 7, 29]. On one hand, social actions of a user reflecting the user's interests are affected by the dynamic nature of social structure. On the other hand, evolution of users' interests can potentially modify the social structure itself [6, 29]. Therefore, the structure of the interaction network, where users perform social actions to keep relationships with each other, is a good indicator for predicting potential interests of users.

Based on the network theory, users in the interaction network can be regarded as nodes, relationships reflected by replying, voting behaviors among users can be regarded as edges between nodes, and the weight of an edge is regarded as the weight of the relationship.

According to social science researchers' research, peer-to-peer influence and homophily effect are two main reasons cause the correlation of users' behaviors[30]. Homophily[31] refers to the principle that "contacts between similar people occur at a higher rate than among dissimilar people", and is vividly summarized as "birds of a feather flock together". Therefore, users with a direct relationship between them are likely to share similar interests. The more interaction two users share, the stronger the correlation intensity is, reflecting users with similar interests will more easily agree with each other. The triadic interaction rule[32] proposed by Heider says that "the friend of my friend is my friend." This theory describes the behavior trend of users affected by the intermediate users, indicating that the users who have an indirect association relationship may have similar interests.

B. System Architecture of the Proposed Model

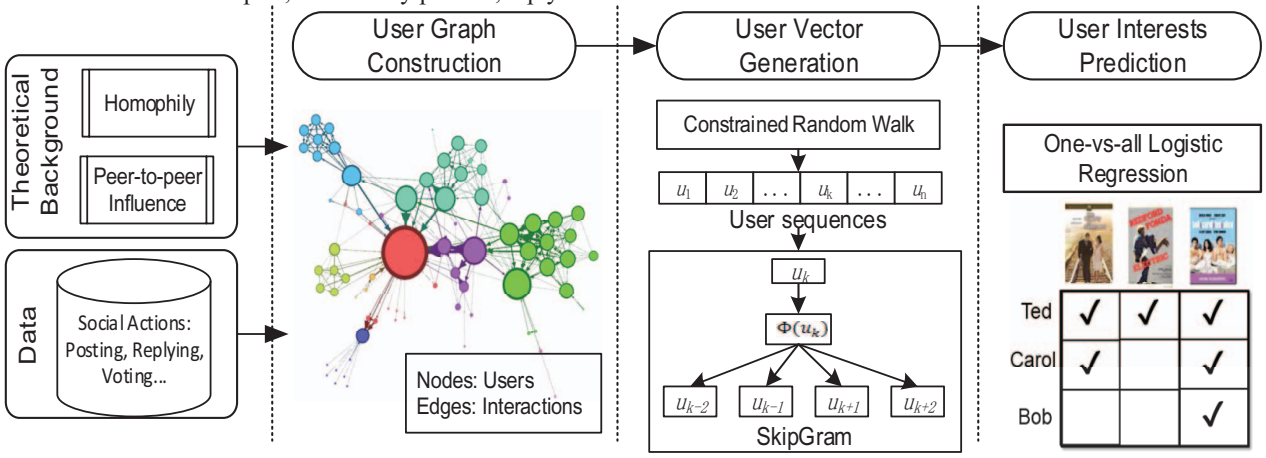

Figure 1. System architecture of the proposed model

The above theoretical background provides theoretical support to our proposed model. The overview of our model is shown in Figure 1, which consists of three functional modules, namely, user interaction graph construction, user latent representation generation, and user interest prediction. In user interaction graph construction module, we obtain user interaction information from social media data, and construct user interaction graph based on social network theory. In user latent representation generation module, we propose a constrained DeepWalk method, which is based on DeepWalk and takes interaction intensity into account, thus help generate user vector representation more exactly. In user interest prediction module, one-vs-all logistic regression method is adopted to predict users' interests.

1) Construction of user interaction network

In social media, information diffuses through users' interaction network, which can be represented as a weighted graph $G = (V, E, W, Y)$, the node set V denotes the user set. A relationship between user u and user v is established when they perform similar social actions on the same object, denoted by an edge $(u, v) \in E$. The weight set $W: E \rightarrow N$ is a function that labels each edge with a weight indicating the frequency of edge (u, v). Y is the label set describing the users' multi-interests.

2) Generation of user latent representation

To capture neighborhood similarity and community membership accurately, a deep learning method is employed to learn user latent representation. Inspired by DeepWalk[27], we model a stream of short constrained random walks to generate a bunch of users sequences and adopt SkipGram model to learn social representation of each user in the sequences. In this way, each user v is encoded in a continuous vector space $\Phi(v) \in R^d$ where d is number of dimensions. We will introduce DeepWalk briefly and then illustrate our proposed model Constrained DeepWalk in detail.

a) DeepWalk

In natural language processing area, neural language models have been widely used to mine the semantic and syntactic of human language[26, 33]. In [33], the authors observed that by using random walks in a graph structure, the frequency where the nodes occur in the walks follows a power-law distribution which is very similar to the distribution of word frequency in natural language. Therefore, DeepWalk employs random walks to generate vertices sequences $(v_1, v_2, \dots, v_n)$. These sequences can be thought of short sentences and phrases in a special language. Inspired from word2vec[18] which is a state-of-art word embedding learning method, in each sequence, DeepWalk aims to maximize the probability of the surrounding nodes given the current node representation:

$$\text{maximize } log\Pr(\{v_{i-w}, \dots, v_{i-1}, v_{i+1}, \dots, v_{i+w}\} | \Phi(v_i)) \quad (1)$$

where w is the size of the window. By integrating neural language models into a graph structure, vertices which have similar neighborhoods will obtain similar representations. This characteristic can be well applied to capture similar users in user interaction graph.

b) Constrained DeepWalk

In social media platforms, the correlation intensity between users is an important factor that reflects user's explicit and implicit interests. The more the users interact with each other, the closer they become and the higher possibility that they share common interests. Therefore, we improve the DeepWalk model and propose a Constrained DeepWalk method which takes users' correlation intensity into consideration. The details of new algorithm are shown below in Algorithm 1:

Algorithm 1 Constrained DeepWalk(G, w, d, γ, t)

Input: graph $G(V, E, W, Y)$
window size w
embedding size d
walks per vertex γ
walk length t
Output: matrix of vertex representations $\Phi \in R^{|V| \times d}$

```
1:  Initialization: Sample Φ from u^{|V|×d}
2:  for i = 0 to γ do
3:    O = shuffle(V)
4:    for each v_i ∈ O do
5:      S_{v_i} = Constrained_RandomWalk(G, v_i, t)
6:      SkipGram(Φ, S_{v_i}, w)
7:    end for
8:  end for
```

Line 2-8 in Algorithm 1 is the main part of our model. The outer and inner loops indicate that we iterate over all vertices in the graph for γ times. All vertices are shuffled at each time to ensure the randomness. In the inner loop, each vertex proceed constrained random walks(line 5) to generate a node sequence and perform SkipGram(line 6) to learn the user latent representations.

- **Constrained Random Walks**

The constrained random walks take the graph G, the initial node v_i and the walk length t as inputs and generate a user sequence $S_{v_i} = \{S^1_{v_i}, S^2_{v_i}, \dots S^k_{v_i}, \dots S^t_{v_i}\}$ where $S^1_{v_i}$ indicates the initial node v_i and $S^{k+1}_{v_i}$ is a node randomly chosen from the neighbors of node v_k. Let $\{v_{k1}, v_{k2}, \dots, v_{kj}\}$ denote the neighbors of user v_k and $\{e_{k1}, e_{k2}, \dots, e_{kj}\}$ denote the corresponding edge weights between each neighbor and v_k. We choose the walk target v_{kj} with the probability as following:

$$\Pr(v_{kj}) = \frac{e_{kj}}{\sum e_{k*}} \quad (2)$$

Different from DeepWalk which samples uniformly, the constrained random walk is more likely to choose neighbor with strong correlation intensity as the next walk target. Therefore, the generated user sequences are more meaningful in real scenarios.

- **SkipGram**

In neural language model, SkipGram[33] maximizes the co-occurrence probability among the words within a window in a sentence. In our model, given the representation $\Phi(v_j)$ of each user v_j in the user sequence, we maximize the occurrence probability of neighbor u_k in the window w where $u_k \in S_{v_i}[j - w : j + w]$. The loss function is defined as follows:

$$J(\Phi) = -log\Pr(u_k | \Phi(v_j)) \quad (3)$$

By utilizing stochastic gradient descent(SGD) to optimize the parameters, we acquire the latent representations of all users which will be leveraged to predict users' interests in the next section.

3) Prediction of user interests

The users in the graph are represented in low-dimensional vectors $\Phi \in R^{|V|\times d}$ and the user interests matrix is denoted as $Y \in R^{|V|\times|y|}$, where y is the set of interests labels. Our objective is to learn a hypothesis that maps user representation in Φ to the label set y. We consider the user interests prediction problem as a multi-label classification problem. One-vs-all logistic regression is adopted to train different classifiers for different interest labels. Each classifier c_i regards the samples with label y_i as positive($y_i = 1$) and the others as negative($y_i = 0$). The classifier c_i maximize the likelihood function below:

$$\max \prod_{k=1}^{|V|} \Pr(v_k)^{y_i}\left(1-\Pr(v_k)\right)^{1-y_i} \quad (4)$$

Where $\Pr(v_k) = \frac{\exp(\beta\cdot\Phi(v_k))}{1+\exp(\beta\cdot\Phi(v_k))}$, β is the parameter vector. At the prediction stage, each classifier outputs a probability indicating the user's interest in the corresponding label. We predict the user's interests by the labels with the top-n probabilities.

IV. EXPERIMENTAL RESULTS AND DISCUSSIONS

A. Dateset

We use API to collect 493,994 E-cigarette flavor-related posts on Reddit from 1 January 2011 to 30 June 2015[5]. The data set contains 34051 threads and 458943 replying posts. JuiceDB provides a very interesting review service for e-cigarette juice, we use API to collect juice reviews from June 2013 to November 2015. The dataset contains 14737 reviews for 4813 juice products.

Based on the collected data, we construct user interaction network using replying and reviewing actions. The summary information of the graphs is shown in Table I.

TABLE I. USER INTERACTION GRAPHS USED IN OUR EXPERIMENTS

Graph	JuiceFlavor	JuiceTag	Reddit
|V|	2438	267	5250
|E|	55748	1774	3362112
|y|	9	89	39
Labels of User Interests	Flavor Categories	Flavor Tags	Communities
Source	JuiceDB	JuiceDB	Reddit

- **JuiceFlavor**: This dataset is a user interaction network constructed from JuiceDB. When two users rate the same e-juice product with high score, a relationship is established between them. The intensity of the relationship is denoted by the number of common products commented by these two users. Nine flavor categories such as sweet, fruity, rich, creamy, etc. represent the coarse user interests in flavor. The weights of 7.6 percentage of the edges in the graph are greater than 1.
- **JuiceTag**: Flavor tags such as strawberry, vanilla, blueberry, etc. provided by users describe users' fine-grained interest in flavor. To gain better insights into users flavor interest, We build a user interaction network which uses 89 flavor tags as labels and the weights of 16.5 percentage of the edges are greater than 1.
- **Reddit**: This dataset is a user interaction network describing users' e-cigarette use preference, which can be represented by 39 communities users have joined in, such as DIY_eJuice, ecig_vendors, Vaping etc. When two users reply the same message, a relationship is established between them. The intensity of the relationship is denoted by the number of common messages replied by these two users. The percentage of edges whose weight is greater than 1 reaches 41.9%.

B. Baseline Methods

- Deepwalk[27]: Deepwalk is recently proposed to learn the latent social representation.
- SpectralClustering[34]: In this method, d-smallest eigenvectors of $\tilde{L}$,the normalized graph Laplacian of G are used to generate a representation in R^d, which will be employed for classification.
- Majority**:** This method uses the labels which most frequently appear in the training set as the users' interest.

C. Evaluation Measures

We now evaluate the performance of the proposed model. The confusion matrix for each label y_i is depicted in Table II.

TABLE II. CONFUSION MATRIX FOR ONE LABEL

		Actual Label	
		y_i	not y_i
Predicted	y_i	tp_i	fp_i
	not y_i	fn_i	tn_i

Accordingly, precision P_i and recall R_i are computed as follows:

$$P_i = \frac{tp_i}{tp_i+fp_i},\ R_i = \frac{tp_i}{tp_i+fn_i} \quad (5)$$

For multi-label classification, we use $macro\_F1$ and $micro\_F1$ [27] to evaluate the results. Firstly, *marco_P*, *macro_R*, *micro_P*, *micro_R* are computed as follows:

$$macro\_P = \frac{1}{q}\sum_{i=1}^{q} P_i \quad (6)$$

$$macro\_R = \frac{1}{q}\sum_{i=1}^{q} R_i \quad (7)$$

$$micro\_P = \frac{\sum_{i=1}^{q} tp_i}{\sum_{i=1}^{q} tp_i + \sum_{i=1}^{q} fp_i} \quad (8)$$

$$micro\_R = \frac{\sum_{i=1}^{q} tp_i}{\sum_{i=1}^{q} tp_i + \sum_{i=1}^{q} fn_i} \quad (9)$$

Where q denotes the number of labels. Hence,

$$macro\_F1 = \frac{2macro\_P \cdot macro\_R}{macro\_P + macro\_R} \quad (10)$$

$$micro\_F1 = \frac{2micro\_P \cdot micro\_R}{micro\_P + micro\_R} \quad (11)$$

D. Experimental Analysis

In this section we present the detailed experimental results, and analyze them thoroughly. Based on the methods used in [27], we randomly sample a portion of data as training set while using the other portion as testing set. And the percentage of training data set ranges from 10% to 90%.What's more, we carry out this procedure 10 times and present the average results in terms of both $macro\_F1$ and $micro\_F1$.

TABLE III. COMPARATIVE PERFORMANCE IN JUICEFLAVOR

	%Labeled Nodes	10%	20%	30%	40%	50%	60%	70%	80%	90%
$macro\_F1$	Constrained DeepWalk	**0.5168**	**0.5327**	**0.5430**	**0.5475**	**0.5517**	**0.5513**	**0.5535**	**0.5462**	**0.5489**
	DeepWalk	0.5056	0.5222	0.5318	0.5370	0.5456	0.5447	0.5362	0.5352	0.5266
	SpectralClustering	0.4507	0.4504	0.4493	0.4485	0.4498	0.4499	0.4510	0.4561	0.4542
	Majority	0.4502	0.4493	0.4511	0.4495	0.4511	0.4495	0.4477	0.4462	0.4460
$micro\_F1$	Constrained DeepWalk	0.7070	0.7245	0.7338	0.7409	**0.7477**	**0.7477**	**0.7493**	**0.7493**	**0.7493**
	DeepWalk	0.7074	**0.7263**	**0.7374**	**0.7418**	0.7475	0.7477	0.7455	0.7466	0.7467
	SpectralClustering	0.7192	0.7183	0.7158	0.7138	0.7151	0.7142	0.7121	0.7174	0.7171
	Majority	**0.7216**	0.7204	0.7213	0.7192	0.7202	0.7188	0.7173	0.7164	0.7134

TABLE IV. COMPARATIVE PERFORMANCE IN JUICETAG

	% Labeled Nodes	10%	20%	30%	40%	50%	60%	70%	80%	90%
$macro\_F1$	Constrained DeepWalk	**0.2111**	**0.2401**	0.2524	**0.2664**	**0.2698**	**0.2713**	0.2757	0.2878	**0.3018**
	DeepWalk	0.2052	0.2397	**0.2593**	0.2606	0.2653	0.2706	**0.2938**	**0.294**	0.2370
	SpectralClustering	0.2035	0.226	0.2322	0.2394	0.2541	0.2485	0.2441	0.2502	0.2820
	Majority	0.2104	0.217	0.2285	0.2237	0.2171	0.2293	0.1982	0.1860	0.1497
$micro\_F1$	Constrained DeepWalk	0.3605	0.3900	0.4071	**0. 4226**	**0.4314**	**0.4423**	**0.4489**	**0.4571**	**0.4641**
	DeepWalk	0.3650	0.3956	**0.4113**	0. 4141	0.4228	0.4269	0.4454	0.4417	0.3930
	SpectralClustering	0.3703	**0.3972**	0.4079	0.4105	0.4218	0.4239	0.4224	0.4274	0.4512
	Majority	**0.3739**	0.3927	0.4048	0.4040	0.3988	0.4064	0.3858	0.3669	0.3559

TABLE V. COMPARATIVE PERFORMANCE IN REDDIT

	% Labeled Nodes	10%	20%	30%	40%	50%	60%	70%	80%	90%
$macro\_F1$	Constrained DeepWalk	**0.3440**	**0.3602**	**0.3834**	0.3906	0.3987	**0.4173**	**0.4438**	**0.4587**	**0.4739**
	DeepWalk	0.3218	0.3594	0.3749	**0.3909**	**0.3990**	0.4147	0.4287	0.4508	0.4666
	SpectralClustering	0.0843	0.0846	0.0853	0.0865	0.088	0.0901	0.0919	0.0978	0.1037
	Majority	0.0815	0.0833	0.0836	0.0846	0.0856	0.0866	0.0888	0.0935	0.1051
$micro\_F1$	Constrained DeepWalk	0.7219	0.7329	0.7379	0.7420	0.7448	0.7466	0.7478	0.7493	0.7532
	DeepWalk	**0.7235**	**0.7378**	**0.7446**	**0.7483**	**0.7517**	**0.7532**	**0.7548**	**0.7584**	**0.7597**
	SpectralClustering	0.5942	0.5947	0.5949	0.5947	0.5946	0.5965	0.5962	0.5977	0.5958
	Majority	0.5927	0.5948	0.5951	0.5956	0.5961	0.5966	0.5954	0.5954	0.5978

1) Predicting users' flavor interests

JuiceFlavor graph and JuiceTag graph are constructed to predict users' general and detailed interests in flavors, respectively. In the experiments of JuiceFlavor, we present the results for Deepwalk and Constrained Deepwalk with the following parameter settings: representation-size=256,

number-walks=80, window-size=10; the parameter representation-size used for spectral cluster is set 500. Table III shows the detailed results. Bold represents the best performance in each column.

In the experiments of JuiceTag, the parameter settings for DeepWalk and Constrained DeepWalk are as follows: representation-size=128,number-walks=80,window-size=10;the representation-size is set to 100 for spectral cluster. The results are shown in Table IV.

It can be seen from the results shown in JuiceFlavor and JuiceTag, that both DeepWalk and Constrained DeepWalk perform better than traditional methods, especially in terms of $macro\_F1$ .This proves that learning users' latent representation based on DeepWalk is able to accurately predict which kind of flavors the users may enjoy.

The results of Majority method in JuiceFlavor are good because of the existence of popular and general flavor categories like sweet and fruity. However, in JuiceTag network, where users' fine-grained flavor interests distributed more uniformly than that in JuiceFlavor, the performances of DeepWalk-related methods are more stable than that of Majority method.

Compared with DeepWalk, the proposed Constrained DeepWalk method performs better than DeepWalk in many situations. In JuiceFlavor, the Constrained DeepWalk ends with a 2% lead in $macro\_F1$, a 0.3% lead in $micro\_F1$; in JuiceTag, it ends with an impressive 7% lead in both $macro\_F1$ and $micro\_F1$. Constrained DeepWalk achieves better scores in most cases especially in JuiceTag, which proves that the weight of link is important.

2) Predicting users' e-cigarette use interests in Reddit

In the experiment of Reddit, we present results of Constrained DeepWalk and DeepWalk with the following parameter settings: representation-size=128,number-walks=40,window-size=10, the parameter of representation-size is set 500 in spectral cluster method. The detailed results are presented in Table V.

It can be seen from Table V that methods related to DeepWalk perform well. Since there are 5250 nodes, 3362112 edges and 39 labels in the Reddit graph, the increment of labels and network density makes the $macro\_F1$ performances of spectral cluster and majority method worse. On the contrary, both Constrained DeepWalk and DeepWalk get an outstanding 36% lead in $macro\_F1$ when training portion reaches 0.9 due to the full use of links among users. The good performance of DeepWalk related methods reveal its robustness for users' multi-interests prediction.

3) Analogous Users Discovery

We regard users who have similar labels as analogous users. To further validate that Constrained DeepWalk makes good use of the links among users, we randomly sample some seed users from the user interaction network of Reddit and present the labels of 8 users who are closest to the corresponding seed user in terms of cosine distance. The results of analogous users discovery using DeepWalk and Conatrained DeepWalk are shown in Table VI and Table VII. Each user is represented with a user number and a label sequence. Each sequence in the square bracket indicates the community labels the user is interested in.

From Table VI and Table VII, both methods can find analogous users who have at least one similar label with the seed user. And Constrained DeepWalk gets 3 users and 4 users completely same with No.95 user and No.3461 user respectively. However, DeepWalk only gets 2 users completely same with the seed users. Therefore, it's necessary to take the weight of links among users into consideration to map similar users to close vector space and predict users' interests.

TABLE VI. TOP 8 NEAREST USERS OF USER NO.95

seed user: 95 [0, 3]	
DeepWalk	Constrained DeepWalk
64:[**3**]	**36**:[**0, 3**]
72:[**3**]	**239**:[**0, 3**]
36:[**0, 3**]	305:[1, 2, **3**]
91:[**3**]	**80**:[**0, 3**]
113:[**3**]	1127:[**0**]
92:[**3**, 14]	1767:[**0**]
20:[**3**]	367:[2, **3**]
178:[**0, 3**]	282:[**3**]

TABLE VII. TOP 8 NEAREST USERS OF USER NO.3461

seed user: 3461 [24, 38]	
DeepWalk	Constrained DeepWalk
3298:[**24, 38**]	**3353**:[**24, 38**]
2705:[**24**]	**3679**:[**24, 38**]
3055:[**24**]	2780:[**24**]
3442:[**24**]	2705:[**24**]
2903:[**24, 38**]	**2920**:[**24, 38**]
3171:[**24**]	3341:[**24**]
2698:[**24**]	3476:[**24**]
2614:[**24**]	**2736**:[**24, 38**]

It can be seen from the above experimental results that our model can accurately predict users' coarse- and fine-grained flavor interests, preference of using e-cigarette. With this valuable mined information, business can recommend e-juice containing users' favorite flavors to users, which will help business make better marketing policies; meanwhile, management department can keep track of the changes of users' interests and adjust their policies accordingly. For example, based on the users' interests predicted by our model, we can provide a "Starb Hurst" juice to some target users who may be interested in strawberry flavor. The accurate recommendation not only improves users' satisfaction, but also increases the sales of "Starb Hurst", it helps merchants get more benefits.

V. CONCLUSIONS

In this paper, we propose a novel model to predict users' multi-interests in health-related applications, which models a stream of short constrained random walks to generate a bunch of users' sequences and adopts SkipGram model to learn social representation of each user. Experimental results on two real world health-related datasets show the efficacy of the proposed model. This work is a first step towards utilizing

social media including Reddit and JuiceDB to predict user's multiple interests in the health-related topics. The model can integrate other information such as direction information of user interaction, users' other kinds of social relationship information, etc. in the future.

ACKNOWLEDGMENT

This research is supported by the NNSFC projects (No. 61172106, 91546112, 91224008) and Important National Science & Technology Specific Projects (No. 2013ZX10004218).

REFERENCES

[1] P. Yan, H. Chen, and D. Zeng, "Syndromic surveillance systems," *Annual Review of Information Science and Technology,* vol. 42, pp. 425-495, 2008.

[2] F. Wang, K. M. Carley, D. Zeng, and W. Mao, "Social Computing: From Social Informatics to Social Intelligence," *Intelligent Systems, IEEE,* vol. 22, pp. 79-83, 2007.

[3] U. Pavalanathan and M. De Choudhury, "Identity Management and Mental Health Discourse in Social Media," in *Proceedings of the 24th International Conference on World Wide Web Companion*, 2015, pp. 315-321.

[4] M. Arthur, "Reddit: Tracking the 2014 Ebola Outbreak across the World," *Nursing Standard,* vol. 29, pp. 32-32, 2014.

[5] L. Wang, Y. Zhan, Q. Li, D. D. Zeng, S. J. Leischow, and J. Okamoto, "An Examination of Electronic Cigarette Content on Social Media: Analysis of E-Cigarette Flavor Content on Reddit," *International journal of environmental research and public health,* vol. 12, pp. 14916-14935, 2015.

[6] R. S. Burt, "Toward a structural theory of action: network models of social Structure, Perception, and Action," 1982.

[7] D. Ganley and C. Lampe, "The ties that bind: Social network principles in online communities," *Decision Support Systems,* vol. 47, pp. 266-274, 2009.

[8] Y. Liang, X. Zheng, D. D. Zeng, X. Zhou, S. J. Leischow, and W. Chung, "Characterizing Social Interaction in Tobacco-Oriented Social Networks: An Empirical Analysis," *Scientific reports,* vol. 5, 2015.

[9] C. Luo, X. Zheng, D. D. Zeng, and S. Leischow, "Portrayal of electronic cigarettes on YouTube," *BMC public health,* vol. 14, p. 1028, 2014.

[10] D. Fan. (2014). *PG vs VG – All Things You Should Know About E-liquid.* Available: https://www.linkedin.com/pulse/20140610083157-323109215-pg-vs-vg-all-things-you-should-know-about-e-liquid

[11] G. Kostygina, S. A. Glantz, and P. M. Ling, "Tobacco industry use of flavours to recruit new users of little cigars and cigarillos," *Tobacco control,* pp. tobaccocontrol-2014-051830, 2014.

[12] FDA. *What are FDA's Regulations for Flavored Tobacco?* Available: http://www.fda.gov/AboutFDA/Transparency/Basics/ucm208085.htm

[13] J.-F. Etter, "Explaining the effects of electronic cigarettes on craving for tobacco in recent quitters," *Drug and alcohol dependence,* vol. 148, pp. 102-108, 2015.

[14] J. Han and H. Lee, "Characterizing user interest using heterogeneous media," in *Proceedings of the companion publication of the 23rd international conference on World wide web companion*, 2014, pp. 289-290.

[15] T. Lappas, K. Punera, and T. Sarlos, "Mining tags using social endorsement networks," in *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, 2011, pp. 195-204.

[16] P. Bhattacharya, M. B. Zafar, N. Ganguly, S. Ghosh, and K. P. Gummadi, "Inferring user interests in the twitter social network," in *Proceedings of the 8th ACM Conference on Recommender systems*, 2014, pp. 357-360.

[17] M. J. Welch, U. Schonfeld, D. He, and J. Cho, "Topical semantics of twitter links," in *Proceedings of the fourth ACM international conference on Web search and data mining*, 2011, pp. 327-336.

[18] W. He, H. Liu, J. He, S. Tang, and X. Du, "Extracting Interest Tags for Non-famous Users in Social Network," in *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, 2015, pp. 861-870.

[19] J. Wang, W. X. Zhao, Y. He, and X. Li, "Infer user interests via link structure regularization," *ACM Transactions on Intelligent Systems and Technology (TIST),* vol. 5, p. 23, 2014.

[20] H. Bao, Q. Li, S. S. Liao, S. Song, and H. Gao, "A new temporal and social PMF-based method to predict users' interests in micro-blogging," *Decision Support Systems,* vol. 55, pp. 698-709, 2013.

[21] H. Ma, T. C. Zhou, M. R. Lyu, and I. King, "Improving recommender systems by incorporating social contextual information," *ACM Transactions on Information Systems (TOIS),* vol. 29, p. 9, 2011.

[22] M. Jamali and M. Ester, "A matrix factorization technique with trust propagation for recommendation in social networks," in *Proceedings of the fourth ACM conference on Recommender systems*, 2010, pp. 135-142.

[23] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *Pattern Analysis and Machine Intelligence, IEEE Transactions on,* vol. 35, pp. 1798-1828, 2013.

[24] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097-1105.

[25] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing,* vol. 20, pp. 30-42, 2012.

[26] R. Collobert and J. Weston, "A unified architecture for natural language processing: Deep neural networks with multitask learning," in *Proceedings of the 25th international conference on Machine learning*, 2008, pp. 160-167.

[27] B. Perozzi, R. Al-Rfou, and S. Skiena, "Deepwalk: Online learning of social representations," in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2014, pp. 701-710.

[28] S. Song, Q. Li, and X. Zheng, "Detecting popular topics in micro-blogging based on a user interest-based model," in *Neural Networks (IJCNN), The 2012 International Joint Conference on*, 2012, pp. 1-8.

[29] D. Lu, Q. Li, and S. S. Liao, "A graph-based action network framework to identify prestigious members through member's prestige evolution," *Decision Support Systems,* vol. 53, pp. 44-54, 2012.

[30] C. F. Manski, "Identification of endogenous social effects: The reflection problem," *The review of economic studies,* vol. 60, pp. 531-542, 1993.

[31] M. Cha, A. Mislove, and K. P. Gummadi, "A measurement-driven analysis of information propagation in the flickr social network," in *Proceedings of the 18th international conference on World wide web*, 2009, pp. 721-730.

[32] F. Heider, "Attitudes and cognitive organization," *The Journal of psychology,* vol. 21, pp. 107-112, 1946.

[33] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *ICLR'13 Workshop,* 2013.

[34] L. Tang and H. Liu, "Leveraging social media networks for classification," *Data Mining and Knowledge Discovery,* vol. 23, pp. 447-478, 2011.