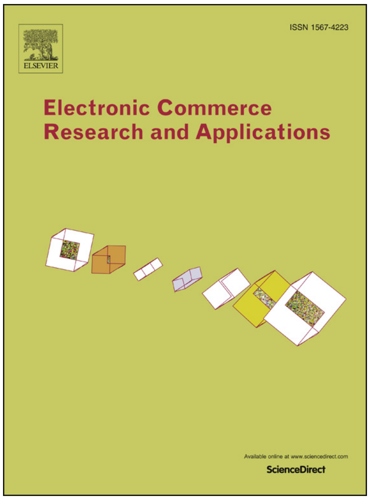
Accepted Manuscript


ISSN 1567-4223
ELSEVIER
Electronic Commerce Research and Applications
Available online at www.sciencedirect.com
ScienceDirect

A graph-oriented model for hierarchical user interest in precision social marketing

Zhiguo Zhu, Yuhe Zhou, Xiaoyi Deng, Xuhui Wang

PII: S1567-4223(19)30022-5
DOI: https://doi.org/10.1016/j.elerap.2019.100845
Article Number: 100845
Reference: ELERAP 100845

To appear in: *Electronic Commerce Research and Applications*

Received Date: 11 September 2018
Revised Date: 3 April 2019
Accepted Date: 3 April 2019

Please cite this article as: Z. Zhu, Y. Zhou, X. Deng, X. Wang, A graph-oriented model for hierarchical user interest in precision social marketing, *Electronic Commerce Research and Applications* (2019), doi: https://doi.org/10.1016/j.elerap.2019.100845

A GRAPH-ORIENTED MODEL FOR HIERARCHICAL USER INTEREST IN PRECISION SOCIAL MARKETING

Zhiguo Zhu[1] (corresponding author), Yuhe Zhou,[1] Xiaoyi Deng,[2,3] Xuhui Wang[4]

[1] School of Management Science and Engineering,
Dongbei University of Finance and Economics, 116025, Dalian, P.R. China
[2] Business School, Huaqiao University, 362021, Quanzhou, P.R. China
[3] Rutgers Business School, Rutgers University, Newark, NJ 07102, USA
[4] School of Business Admin., Dongbei Univ. of Finance and Economics, 116025, Dalian, P.R. China

Last revised: April 3, 2019

ABSTRACT

With the rapid development of social commerce, how to push and diffuse marketing messages in online social network (OSN) more effectively has increasingly become a significant issue, which can result in benefits for enterprises, users and platforms. A fundamental solution to this issue is how to accurately and comprehensively model user interest. To resolve such a significant and challenging task, our study constructed a *user interest graph* represented by a hierarchical tree structure that covers a wide range of topics, from coarse-grained to fine-grained three-level interest topics, such as food, entertainment and shopping, with a total of 167 nodes. In addition, considering that a user's interests are always changing over time, an exponential interest decay scheme is employed in this study. Finally, a series of experiments are conducted to evaluate the performance of the proposed model by comparing it with three benchmarks designed based on the proposed algorithm and two similar hierarchical user interest models. The experimental results demonstrate our model works well to predict user interests. This research will provide important basic technology and valuable decision support for precise and personalized social marketing practices.

Keywords: Feature extraction, precision social marketing, semantic similarity, social commerce, user interest graph

Acknowledgments. The authors greatly appreciate the valuable and insightful comments and suggestions raised by the editors and reviewers, who have certainly been of great help to improve the quality of this article. This work was supported and funded by the National Natural Science Foundation of China (No.71672023, 71401058, 71672026, 71272050, 71771040); the Humanities and Social Sciences Foundation of the Ministry of Education of China (No.16YJAZH083); and the General Research Project Funds of Dongbei University of Finance and Economics (No. DUFE2017Y03).

1. INTRODUCTION

Interpersonal interactions – such as information transmission, emotional communication, business transactions – have been enhanced by emerging and diverse *online social networks* (OSN), such as Facebook, Twitter and Dianping.[1] Gradually, a virtual society emerged, which dominates the bulk of human digital footprints, including relationships and behaviors (Lazer et. 2009, Freeman 2004).

1.1. Research Background

In this context, how to utilize behaviors, such as recommending, reviewing, forwarding and sharing among users in the virtual society (Mislove 2009, Zhu 2013) to carry out effective marketing activities (Li and Shiu 2012), has become one of the most important issues in the social commerce revolution (Steven and Olivier 2010, Han et al. 2018).

The simple and direct pattern of *push-forward-diffusion* marketing messages by leveraging asocial graph has been widely adopted in current practice and academic research on social marketing (Turban et al. 2015, Zhu et al. 2016).). Yet, because of inadequate consideration of user interests and preferences, this pattern can easily result in an uninterested user's antipathy in marketing message diffusion in OSNs. Moreover, this simple pattern’s lower precision will undoubtedly increase the marketing cost of enterprises, while the effects and efficiency often remain unsatisfactory. Further, to improve user experience and enhance users' stickiness, a growing number of social platforms have also begun to restrict the indiscriminate flooding with marketing messages. Therefore, the lack of precision and personalization in current social marketing practices has been a prominent problem that brings trouble to users, enterprises and platforms (Burchell et al. 2013). It is worth noting, though that there actually is a user interest graph in OSN besides the user social graph shown in Fig. 1 as an example.

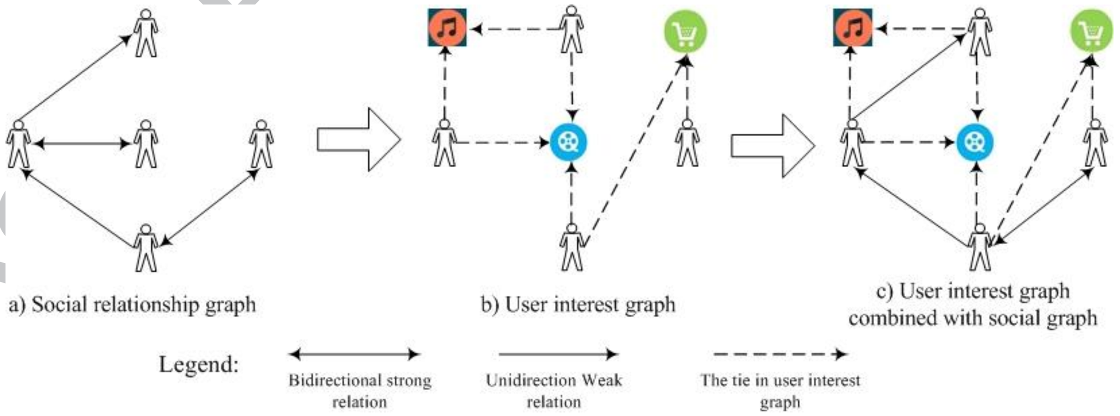

a) Social relationship graph
b) User interest graph
c) User interest graph combined with social graph
Legend:
Bidirectional strong relation
Unidirection Weak relation
The tie in user interest graph

Fig. 1. An example of social relationship graph, user interest graph and their combination

Thus, a key issue for improving accurate, personalized social marketing is how to construct a model of user interest, and then effectively infer user interests from their profile in OSN. This line of research has also attracted extensive attention from research fields such as decision support systems in

[1] Dianping, www.dianping.com, is a leading Chinese online social platform focusing on consumers and third-party consumption reviews.

e-commerce and marketing science (Mayer 2009, Zabin and Brebach 2008).

1.2. Challenges and Main Contributions

For this interesting and significant issue in social marketing, we propose a model of an inverted tree-shaped *user interest graph* (UIG) and its corresponding unsupervised algorithm by means of extracting and mining multidimensional user generated contents (UGC) and interaction records in the user's social network profile. The Chinese OSN Dianping was chosen for this work. We extracted the feature terms from items (in "collections," "reviews," etc. of the user profiles on the site). Next, our proposed algorithm generates predictive scores on all of nodes in the tree to reflect the distribution and extent of user interest. Our algorithm results will provide basic technology and valuable decision support for more precise and personalized social marketing practices.

It is difficult to construct such a model for predicting user interests based on OSN. The main challenges we face include:

(1) The UGC and interaction information extracted from user profiles are often multidimensional and heterogeneous. They also are always noisy and unnormalized. Hence, how to preprocess these data separately is an issue that needs to be addressed.

(2) In a social network profile, the topics a user is interested in are not always directly expressed. In addition, the item categories listed in user profile in OSN are often of coarse granularity, flat and inconsistent with the current general interest categories. Thus, it is a challenging problem to build a tree-like structure of UIG with the item category and then accurately predict the interest scores.

(3) The various items recorded in the user profile tend to change over time, so how to calculate and reflect the evolution of user interest in our proposed model is a difficult issue.

The main contributions of this research are:

(1) A three-level structure of UIG is constructed, and the semantic similarity is calculated between the feature terms extracted from the items in a user's profile and the interest nodes of the UIG one by one. This way, the user's explicit and implicit interests from coarse-grained to fine-grained interest topics can be inferred.

(2) The dynamic nature of user interest is fully considered in our model. Therefore, predicting results can more accurately to reflect the on-the-ground truth.

(3) We conduct a series of elaborate experiments to compare our predictive interest values with user's real interests obtained from investigation. The final experimental results demonstrate that the performances of this model on some typical metrics are advantageous.

2. LITERATURE REVIEW

We offer a brief review and discussion of existing research work in the field of user interest modeling. By comparing the existing work, we further highlight the contributions of this research.

2.1. Explicit Interest Acquisition and Implicit Interest Inference

In the field of user interest modeling, there are two ways to obtain user interest: explicit interest acquisition and implicit interest inference (Hanani 2001). Explicit interest acquisition is carried out mainly by means of asking users to directly input or provide feedback (e.g., evaluating resources, or adding tags) (Carmagnola et al. 2007). For example, the Pocket Restaurant Finder system (McCarthy 2002) can directly obtain every customer's dining preferences according to their score (e.g., distance range, expenditure, and environment). PolyLens directly collects user preferences through requiring users to rate movies (O'Connor et al. 2002). By collecting user scores on different items in a website, a utility matrix representing the user's interests can be built. The rows represent the user, and the columns represent each user's interest scores (Rajamaran and Ullman 2011).

Schafer et al. (2001). constructed recommendations by collecting feedback from readers about books they read. Wen et al. (2012) also constructed a personalized news recommender system based on user-defined reading preferences. However, when taking part in these activities, sometimes users do not have a positive attitude. Also, a relatively complete list of interests cannot be provided by many users, since their interests are often distributed across different environments. On the other hand, implicit interest inference utilizes such information as browsing behavior, generated contents, etc., instead of information on preferences directly provided by the user. Also such user preference data will change over time and can be detected (Facca and Lanza 2005). The research on inferring a user's implicit interest has gradually become more mainstream, and has emerged as a hot topic in this field. Moreover, the methods for inferring implicit interest can further be divided into the following two subfields, which we will discuss in the next two subsections.

2.2. Modeling User Interest from Web Server Logs

In the Web 1.0 environment, user interest modeling mainly belonged to the web usage mining area. (Facca and Lanza 2005). This method extracts the features from the access logs stored in servers, including browsing behavior (e.g., "duration") (Liang and Lai 2002, Raphaeli et al. 2017), browsing content (Seo and Zhang 2001) (e.g., "viewed web pages"), and clickstream data (Su and Chen 2015) to infer and mine user interests. Earlier studies monitored user browsing behavior. Sakagami and Kamba (1997), Pazzani et al. (1997), Lieberman (1995), and Linden et al. (1997) attempted to build user interest models based on the user providing implicit and explicit interaction information with the system, and then predicted and recommended the user web pages of possible interest. Qiu and Cho (2006) researched how a user's interest can be automatically inferred based on her past click history and further be used to generalize the personalized search results. Claypool et al, (1997) and White et al. (2009) explored the correlation of implicit indicators and explicit interest respectively, and the predictive ability of different background information sources about user interest. Chan (1999) proposed a model consisting two parts: WAG and PIE. WAG records the users' web access patterns, and PIE learns users' interests based on page contents, and reorders the results of the search engine. Akcayol et al. (2018) proposed a new *weighted multi-attribute-based recommender system* (WMARS) developed using extended user behavior analysis including: the number of clicked items in the recommendation list, duration of tracking, likes/dislikes, association rules of clicked items, etc. To sum up, in the context of Web 1.0, due to the limitations of a single information source and the scarcity of information, the predicted results by these methods are often inaccurate and incomplete.

2.3. Modeling User Interest from OSN

In the Web 2.0 context, the user profile in OSN can provide richer and multi-dimensional information, including user generated contents (e.g., product review, recommendation), user individual behavior (e.g. tagging, collecting) (Zhu et al. 2015) and interactions among users (e.g., forwarding, recommending (Hogg 2010). Capturing and understanding user interests clearly are an important part of social media analytics, which is getting increased attention.

The research on inferring interest from user generated contents is well developed. Bao et al. (2013) constructed a temporal and social probabilistic matrix factorization model to predict potential user interests in micro-blogging. Asur and Huberman (2010) discovered that box-office revenues of movies can be successfully predicted by analyzing users' interest in micro-blogging. Banerjee et al. (2009) gathered tweet data across ten cities worldwide for a period of four weeks to generate a list of keywords. They then employed mining and statistical methods to discover the distribution of user interests for categories such as "games," "food" and "movies."

Xu et al. (2010) proposed an improved author-topic model to infer user topics of interest on Twitter by filtering out interest-unrelated tweets from the aggregated user profiles. A new collaborative filtering recommender system was introduced by Nguyen (2017), which is offered a new methodology: soft ratings. They can be used for modeling subjective, qualitative, and imperfect information about user preferences, and for a more realistic and flexible means for users to express their preferences on products and services. Kapanipathi et al. (2014) exploited the hierarchical semantics of concepts from tweets to infer richer user interests expressed as a hierarchical interest graph. This relates to semantic similarity calculation between items and interest topics in our model.

From the perspective of user individual behavior and interaction, Abel et al. (2011) explored whether a user's professional scientific interests overlap with his social network interactions and can be used to recommend relevant publications. Ying et al. (2018) provided new insights into user activity in today's OSNs, in particular the posting frequency and temporal patterns, and suggested a framework for profiling users based on their posting activities. Ho et al. (2012) estimated the users' shared interests based on whether users liked the Facebook pages for four popular interests, and then studied how the shared interests influenced conversations and friendships on Facebook. The tags of interest on web resources are first-hand information directly given by users without any middleman modification. Goel et al. (2018) leveraged the concept of semantic relatedness for tag clustering to construct a strong *user interest profile* (UIP), which provides a complete list of user preference along with his area of interest. However, other important UGC tags for inferring user interest, such as comments and reviews, were neglected in this work. In our model, a richer source of items will be extracted from four representative sections for online user profile to infer user interests.

The research on user interest modeling by mining user profile includes: Karatay and Karagoz (2015) proposed a *named entity recognition* (NER) model for Twitter user interests based on user profile modeling. Garcia Esparza et al. (2013) presented a user profiling model based on topical categorization of URLs in tweets. In their work, a mean profile prediction accuracy of 0.73 for 32 users over 18 coarse-grained interest categories was achieved. Zheng et al. (2019) developed a *hierarchical interest overlapping community* (HIOC) *detection method* by studying similar relationships between user profiles, and further presented a personalized recommendation model. Li et al. (2011) built a hierarchical user interest model labeled with the topic for each cluster, and then proposed a *graph-based Chinese phrases hierarchical clustering algorithm* (GCPHC). It organizes the user interest in a hierarchy tree structure to map user interest to topics. This three-level tree structure of user interest topics is also employed in our study. Ma et al. (2011) predicted users' higher-level interests based on terminology-specific keywords extracted from their profiles on social networks. However, this supervised method has some limitations: (1) keyword extraction is domain-specific; and (2) Pre-defined ontology for each domain is required. In our study, a widely-used Chinese ontology, HowNet (2019), is employed as the basis for sematic similarity calculation.

To sum up, we constructed a novel, general model of UIG evaluated with a real dataset collected from Dianping.com. This model can be easily applied to OSNs in other languages after necessary extensions and modifications. Moreover, the proposed model not only can infer a user's wide interests in terms of 167 topics on three levels. This study also considers the interest decays over time. Finally, the experimental results demonstrate the proposed model has better performance compared with other benchmarks and two similar hierarchical user interest models.

3. THE UIG MODEL

3.1. The Structure of the UIG Design

In this study, the interest nodes in the UIG structure not only can cover the main popular interest topics of users, but also the number of interest topic levels with different granularities need to be appropriate. So if there are too few levels (one or two), the representation of the user's topics of interest will be too rough. In contrast, if too many levels are designed into the structure, it will be too detailed and lead to computational complexity. The three-level tree structure has been widely used in many studies (Su and Chen 2015, Li et al. 2011, Kapanipathi et al. 2014). In addition, many real-world e-commerce websites, such as Dianping, Yelp and Amazon, also present their product categories using a three-level hierarchical structure. This study constructs an inverted tree-shaped structure containing three level of popular interest nodes to represent the user interest graph. This is based on the page categories of Dianping.com and Yelp.com, which cover a wide range of popular Chinese and American lifestyle areas and interests. Further, some of the very fine-grained (too specific) and similar interest topics have been filtered or merged in the process of tree construction.

Eventually, the three-level inverted tree-shaped structure contains a total of 167 nodes. As shown in Fig. 2, the distribution of nodes in the tree includes: a root node denoting a user; 6 nodes on top level (Lv1); 39 more specific interest nodes on the level two (Lv2) and 122 fine-grained nodes on the level three (Lv3). So note that the figure just shows a partial view of the UIG tree.

A list of notation in the equations in this section of the article is in Table 1.

Table 1. Modeling Notation

Name	Definition	Source
U^{IG}	A user's interest graph.	Eq. 1
I^{u}	A set of items recorded in the profile of user *u*.	Eq. 2
I^{u}_Ft	Set of feature terms *Ft* extracted from item set I^{u}.	Eq. 3
Ft_{i_k}	Set of feature terms extracted from item i_k.	Eq. 3
$\omega_{I^u}(i_k)$	Weight of item i_k in I^{u}.	Eq. 4
SS_j	Value of semantic similarity between feature term *j* and interest node *n*.	Eq. 5
$ASS_{i_k}^{n}$	Value of average semantic similarity between item i_k and *n*.	Eq. 6
$WSS_{i_k}^{n}$	Value of weighted semantic similarity between item i_k and *n*.	Eq. 7
TW_n	Timed weight of interest node *n*.	Eq. 8
$<WSS_{i_k}^{n}, T_{i_k}^{last}>$	With two members: $WSS_{i_k}^{n}$ and the timestamp $T_{i_k}^{last}$ of generated or updated item i_k.	Eq. 8
CW_n	The cumulative weight of node *n*.	Eq. 9

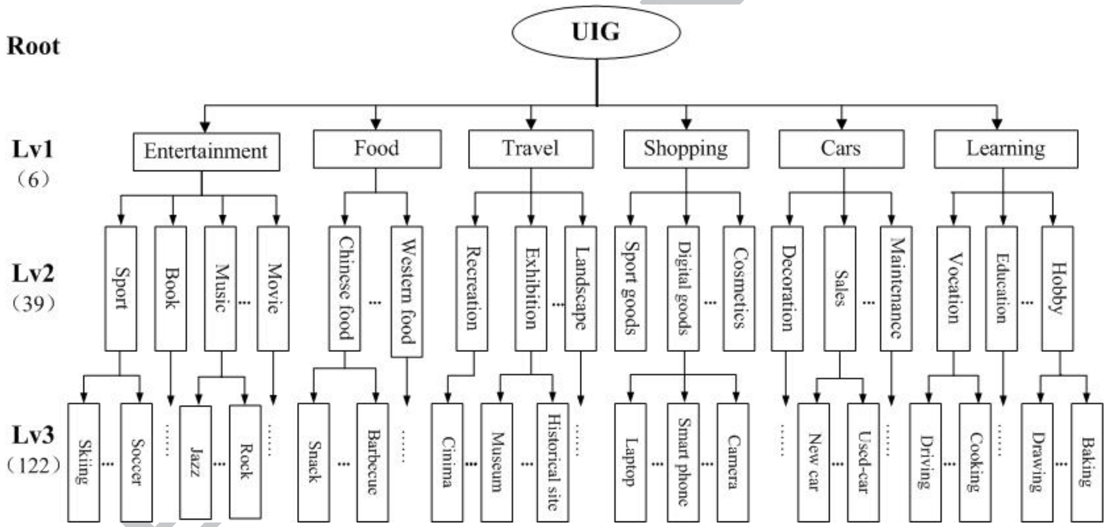


Fig. 2. The designed structure of user interest graph (partially showing)

According to the structure of UIG, a formal definition is:

$$U^{IG} = \left\{ n_1 \middle| n_{11}(n_{111}, n_{112}, \ldots), n_{12}(n_{121}, n_{122}, \ldots), \ldots \middle| , \ldots, n_i \middle| n_{i1}(n_{i11}, n_{i12}, \ldots), n_{i2}(n_{i21}, n_{i22}, \ldots), \ldots \middle| \right\} \quad (1)$$

where U^{IG} represents a user *U's* interest graph and *n* denotes an interest node in the tree, whose subscript corresponds to its hierarchical sequence number. For example, n_{i21} represents the first leaf node on the third level, which belongs to the second leaf node under the i^{th} top node.

3.2. The Main Steps of Calculating the Interest Scores in the UIG Tree

As we know, important items restored in the online user profile can be used as treasures for explicitly or implicitly expressing user interest. Therefore, the items first need to be collected from some representative sections of Dianping user profile in our study. Next, based on the structure of UIG we built, the semantic similarity can be calculated between feature terms extracted from the collected items and interest nodes in the tree one by one, and then the weight of every item can also be calculated. Finally, we predict the interest scores so as to accurately quantify the extent to which user is interested in every interest node in the tree. It is worth pointing out that if a certain interest topic cannot be matched with any node after a complete walking in the tree, this topic will be finally added into this user's UIG tree as a newly-discovered interest node.

Accordingly, we propose the UIG model framework as shown in Fig. 3. The main steps of the four modules in the framework are discussed next.

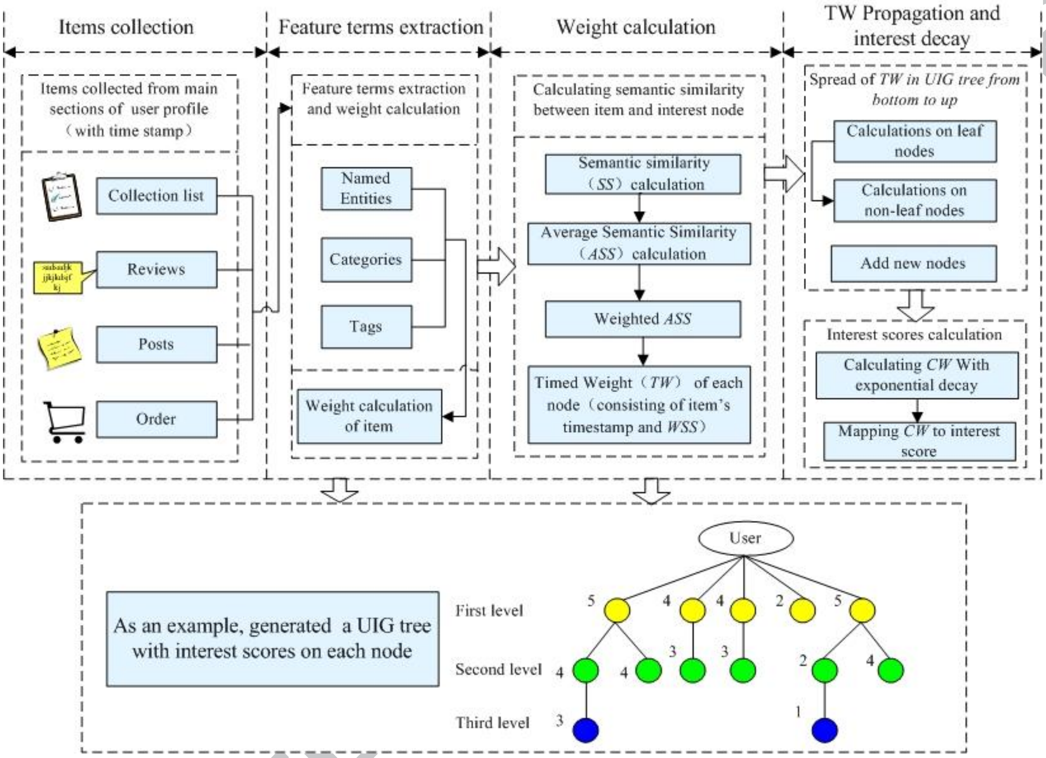

Items collection
Feature terms extraction
Weight calculation
TW Propagation and interest decay
Items collected from main sections of user profile (with time stamp)
Collection list
Reviews
Posts
Order
Feature terms extraction and weight calculation
Named Entities
Categories
Tags
Weight calculation of item
Calculating semantic similarity between item and interest node
Semantic similarity (SS) calculation
Average Semantic Similarity (ASS) calculation
Weighted ASS
Timed Weight (TW) of each node (consisting of item's timestamp and WSS)
Spread of TW in UIG tree from bottom to up
Calculations on leaf nodes
Calculations on non-leaf nodes
Add new nodes
Interest scores calculation
Calculating CW With exponential decay
Mapping CW to interest score
As an example, generated a UIG tree with interest scores on each node
User
First level
Second level
Third level
5
4
4
2
5
4
4
3
3
2
4
3
1

Fig. 3. The framework of the UIG model

3.2.1. Module 1: Item Collection

In Module1, four representative sections of "Mine" (representing the user profile in Dianping), including "collections," "posts," "reviews," and "orders," are chosen for the collection of items in this study. The interface of four selected sections in Dianping is shown in Fig. 4.

Although the items are recorded in different sections respectively, they may belong to a same or similar topic. In Eq. 2, I^u is defined as the set of items recorded in the profile of a user *u*.

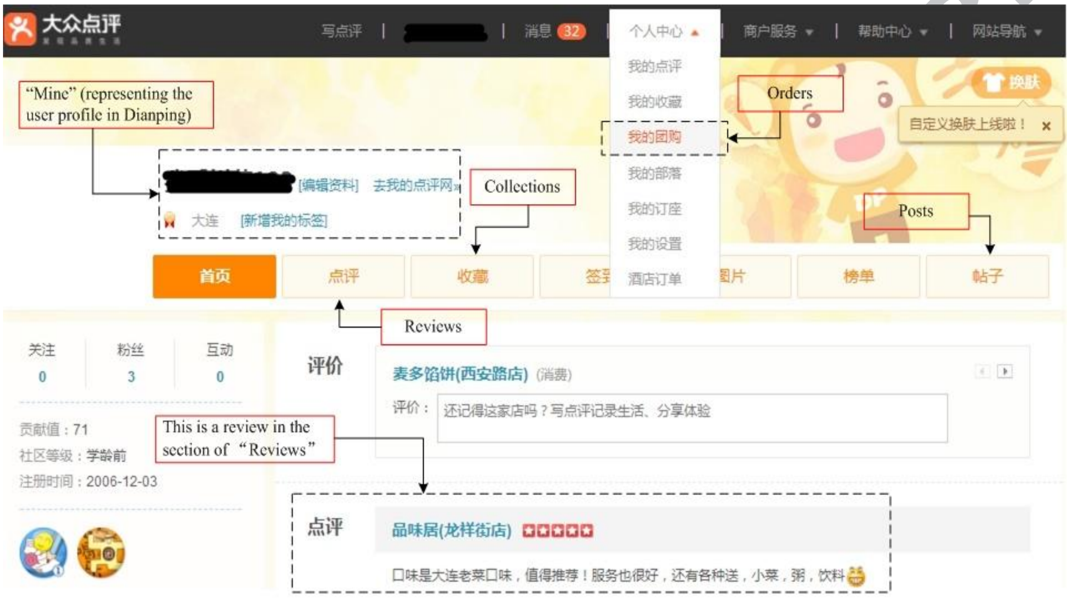
$$I^u = \{i_1, i_2, ..., i_k, ..., i_m\} \tag{2}$$

where i_k represents the k^{th} recorded item and *m* represents the total number of items.

In the set of I^u, the fine-grained feature terms will be extracted from each item, so as to accurately and comprehensively infer user interest scores on every node of UIG.

3.2.2. Module 2: Feature Term Extraction

For comprehensively and deeply inferring user interests, the items in I^u are still too coarse-grained. Hence, more fine-grained feature terms with rich semantic information need to be further extracted. For this, we employ NLPIR2016 in this study, a Chinese software tool for text mining and analysis. Through some preprocessing steps of NLPIR2016, Chinese texts in all of the items of I^u can be segmented, and then the finer-grained feature terms (Named Entities, Tags and Categories) can be extracted and identified. This study is only intended to estimate user interests rather than dislikes, so the feature terms we obtain have to be subjected to sentiment analysis by using NLPIR2016. After sentiment judgment is done, only positive or neutral feature terms will be retained while the negative ones will be eliminated. Moreover, taking into account the dynamically changing user interests over time, the timestamp of each item (indicating when it was generated or last updated) needs to be stored as a label information for further utilization in the subsequent steps.



还记得这家店吗？写点评记录生活、分享体验

口味是大连老菜口味，值得推荐！服务也很好，还有各种送，小菜，粥，饮料

Figure 4. A screenshot of the interface of Dianping for the user profile "Mine"

For example, a user has a collected commodity page (Huawei mate9 mobile phone) in his user profile. From this page, some important feature terms can be extracted:

- **Named Entities:** Smart phone (digital products), Huawei (brand) Made in China (country).
- **Categories:** Shopping.
- **Tags:** Digital product, Communication tool, 5/23/2017 (generated timestamp).

From these extracted feature terms we can further calculate the user interest scores on the more fine-grained interest topics, such as "digital products" (a second-level node under the "shopping" node in the tree of UIG) and "smart phone" (a third-level node). Thus, the extracted feature terms from I^u provide valuable corpus for inferring fine-grained user interests with accuracy.

Finally, the set of feature terms Ft extracted from item set I^u can be represented as I^u_Ft, in:

$$I^u\_Ft = \{(i_1, Ft_{i_1}),\ldots,(i_k, Ft_{i_k}),\ldots,(i_m, Ft_{i_m})\} \tag{3}$$

where Ft_{i_k} is the set of feature terms extracted from item i_k ($i_k \in I^u$).

3.2.3. Module 3: Weight Calculation

The weight calculation consists of:

(a) Computing the weight of an item in I^u. For each item in I^u_Ft , a normalized Item Weight w_i should be calculated, which represents its contribution for representing user’s interest. For example, assume there are 50 items in I^u, in which 30 items are about “smart phone.” It can be inferred that users are interested in the topic,“smart phone.” So it is reasonable that the item “smart phone” is assigned a higher weight. The Item Weight is calculated as:

$$\omega_{I^u}(i_k) = \frac{\left(\dfrac{tf(i_k)}{tf_{\max}}\right)}{\sqrt{\sum_{k=1}^{\left|I^u\right|}\left(\dfrac{tf(i_k)}{tf_{\max}}\right)^2}}, \quad i_k \in I^u \qquad (4)$$

where $\omega_{I^u}(i_k)$ and $tf(i_k)$ represent the weight and frequency of i_k in I^u , respectively. tf_{max} is the maximum frequency in I^u and $\left|I^u\right|$ is the total number of items.

(b) Calculating semantic similarity between feature terms and interest nodes. In this step, the semantic similarity calculation between *Ft* (the set of the feature term of each item) and interest nodes in the tree is made. Its purpose is inferring the extent of implicit interests, which the user has not expressed in his profile. For example, there is a high semantic similarity between "footballs" and "sporting goods": the similarity value is 0.7). If a user expresses a strong interest about "football" in some items (e.g., collecting football match pages, or making a lot of comments about football news), he is highly likely to go “shopping” and buy some sporting goods related to “football.” The items "football" and "sporting goods" are actually nodes in different interest categories (the former is "entertainment" while the latter is "shopping." (See Fig. 2.) Thus, the implicit interests on the nodes of "shopping" and “sporting goods” can be inferred by means of the sematic similarity calculation.

Based on this idea, the values of all nodes on the tree are initially set at 0, and then the *semantic similarity* (*SS*) value between each feature term in *Ft* and the interest node is calculated. In this study, for Chinese text in “Mine” in Dianping, a *semantic textual similarity method* (Li and Li 2001) is called to compute the *SS* scores. It utilizes the tree hierarchical structure of "sememe" in HowNet (1019) (an ontology, as the basis for *SS* measurement) to calculate the similarity of "sememe," and then the similarity between words (as a set of sememes) can be obtained. Thus, *SS* is found via Eq. 5:

$$SS(W_1,W_2)=max_{i=1\,\ldots,n;j=1,\ldots,m}\,Sim\left\{(S_{11},\ldots,S_{1i},\ldots,S_{1n})\middle|(S_{21},\ldots,S_{2j},\ldots,S_{2m})\right\} \qquad (5)$$

$$in\ which\ \ Sim(S_{1i},\ S_{2j})=\frac{\alpha}{\alpha+distance(S_{1i},\ S_{2j})}$$

where $(S_{11},\ldots,S_{1i},\ldots,S_{1n})$ is the sememe set for word1, with W_1 *and* S_{1i} denoting the i^{th} sememe. In the same way, S_{2j} is the j^{th} sememe in the set for word2 W_2. $Distance(S_{1i},\ S_{2j})$ represents the path length in the tree structure of Hownet between two sememes, S_{1i} *and* S_{2j}, and α is an adjustment parameter.

At the same time, for some English text in Dianping’s user profiles, also adopted is the Stanford

NLP (2019) is as a supplementary tool. Finally, the *SS* is calculated as a cosine similarity value, whose range is between [0, 1] (0 no similarity, 1 complete similarity). After obtaining *SS* values, the *average semantic similarity* (*ASS*) of this node can be calculated by using Eq. 6.

$$ASS_{i_k}^{n} = \frac{\sum_{j=1}^{|Ft_{i_k}|} SS_j}{|Ft_{i_k}|} \tag{6}$$

Here $ASS_{i_k}^{n}$ represents the average value of *SS* between item i_k and interest node *n*; and SS_k represents the value of semantic similarity between feature term *j* in the set of $|Ft_{i_k}|$ and node *n*.

Eq. 6 shows that *ASS* actually reflects the overall semantic similarity between an item and an interest node in the tree. To simplify the problem, the *ASS* values below a given threshold $SS_{threshold}$ will be set as 0, and will not be used in later calculations. The purpose is to eliminate the noise and false positives. Due to cosine similarity measurement of *SS,* the threshold $SS_{threshold}$ is set as the widely-accepted value of 0.293 (which equals 1 - cos (π/4)).

Next, the *weighted semantic similarity* (*WSS*) can be acquired after *ASS* is multiplied by the normalized weight of item:

$$WSS_{i_k}^{n} = ASS_{i_k}^{n} \times \omega_{I^u}(i_k) \tag{7}$$

Here $WSS_{i_k}^{n}$ represents the weighted semantic similarity between item i_k and interest node *n*.

(c) **Computing the timed weight (*TW*) for each node.** The longer an item has been generated or updated, the more user interest about it tends to decay. Due to the dynamic evolution of user interest, an item's generated or updated timestamp will be added in the calculation of *WSS*. As a result, the timestamp *T* of each item and the *WSS* are combined to form a pair for representing a *timed weight* (*TW*) of an interest node. Through iterative calculations, each item pair will be recorded in the *TW* set, as shown by Eq. 8:

$$TW_n = \left\{ \left\langle WSS_{i_1}^{n}, T_{i_1}^{last} \right\rangle, \ldots, \left\langle WSS_{i_k}^{n}, T_{i_k}^{last} \right\rangle, \ldots, \left\langle WSS_{i_m}^{n}, T_{i_m}^{last} \right\rangle \right\} \tag{8}$$

where TW_n represents the timed weight of interest node *n,* the pair $< WSS_{i_k}^{n}, T_{i_k}^{last} >$ contains two members: $WSS_{i_k}^{n}$ and the timestamp $T_{i_k}^{last}$ (generated or last updated item i_k.).

3.2.4. Module 4: *TW* propagation and Interest Decay

(a) Spreading the *TW* from bottom to up in the tree. In the proposed UIG model, the process of spreading the *TW* from the leaf nodes on lower level to their parent nodes on higher level is critical. The reason is that the upward spread process ensures that high-level interests can be accurately inferred from low-level fine-grained interests, which are often expressed by the various extracted feature terms. The spread process of *TW* can be explained by example, as shown in Fig. 5, as follows:

(1) In Step 1 of Fig. 5, it is assumed that each node on the three levels of a UIG tree has

already obtained its set of *TW*.

(2) In Step 2, add the pair $<WSS^{n}_{i_k}, T^{last}_{i_k}>$ of current item i_k to the *TW* set of node n_{111} as the k^{th} weight, which can be expressed as: $TW_{n_{111},\ k} \Leftarrow \left\langle WSS_{i_k}, T^{last}_{i_k} \right\rangle$.

(3) The Step 3 is upward spreading $TW_{n_{111},w}$ to its parent node. For the leaf nodes on the bottom, after comparing the *TW* values among the sibling nodes, the maximum value is added into the existing *TW* set of their parent node. The formal description is:

$$TW_{n\_parent,k} = <Ti^{last}, \textit{arg max}(TW_{n,k} \textit{ and sibling nodes})>$$

For example, after comparing the weights of n_{111} and n_{112} (marked with yellow), it is assumed that the bigger value $TW_{n_{111},k}$ spreads to the *TW* set of its parent node n_{11} :$TW_{n_{11},k'} \Leftarrow TW_{n_{111},k}$.

(4) In the spreading, for the non-leaf nodes the average value of sibling nodes is added to their parent's *TW* set. The formal description is:

$$TW_{n\_parent,k'} = <Ti^{last}, \textit{Average}\ (TW_{n,k} \textit{ and sbiling nodes}) >.$$

As shown in Fig. 5, in Step 4 the weights of n_{11} and its sibling node n_{12} (in green) are averaged, and the value obtained is added to the existing *TW* set of their parent node n_1 (in blue): $TW_{n1,\ k'} \Leftarrow Avg(TW_{n_{11},k'}, TW_{n_{12}})$.

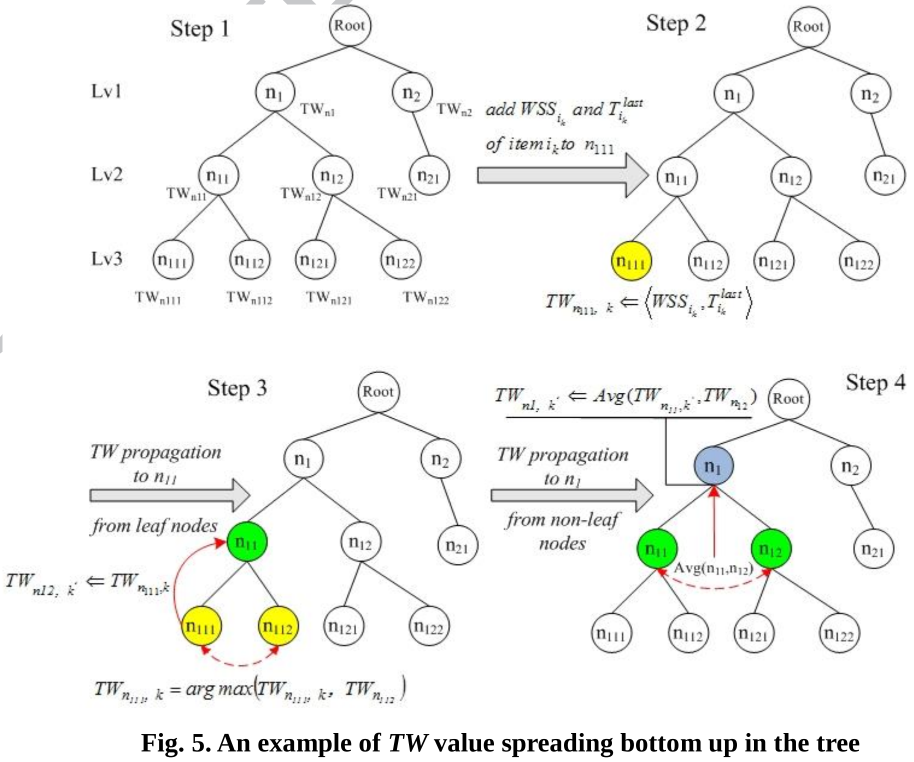


Fig. 5. An example of *TW* value spreading bottom up in the tree

(b) **Discovering and adding a new interest node.** If the maximum *ASS* value between every *Ft* and interest node on three levels is lower than the $SS_{threshold}$, it means that this item is unrelated to any node. In this case, this item will be added as a newly-discovered topic node of interest in the UIG tree of that user.

(c) **Calculating the cumulative weight of each node in UIG tree.** In this process, the Cumulative Weight (*CW*) needs to be calculated after the iterative process of bottom-up spreading, *TW*, concludes. Considering the time decay of interests, the *CW* is defined and calculated with exponential interest decay using the timestamp recorded in the set of *TW*. The calculation formula is:

$$CW_n = \prod_{i=1}^{|TW_n|}\left(1 - TW_{n,i} \times exp^{-\lambda\left(T^{pre} - T_i^{last}\right)}\right) \qquad (9)$$

where CW_n is the cumulative weight of node *n*; and $TW_{n,i}$ is the i^{th} *TW* value in the *TW* set of *n*. T^{pre} and T_i^{last} (unit: of time: a day) represent the present and last time when item *i* was updated or generated, respectively. And λ is a set of exponential decay constant.

(**d) Calculating each node's interest score.** To compare the *CW* value with the 1 to 5 Likert scale scores, the *CW* of each node is further mapped to a Likert scale score as a node's interest score by using the inverse cosine function. To be specific, if the *CW* value is in range of $\left[cos0, cos\frac{\pi}{10}\right)$, then it should be mapped to a 5 score ('Most Interested'. Similarly, if the *TW* value is in $\left(cos\frac{2\pi}{5}, cos\frac{\pi}{4}\right]$, the mapped *CW* will be 1 ('No Interest').

3.3. The Algorithm Description for the Proposed UIG Model

We next present an unsupervised algorithm, according to the main calculation steps of this model:

Algorithm Description: Generating a User's UIG with Interest Scores of the Nodes

```
Input: (1) A user u's profile data in a social network platform (including important items in the "collections,"
           "posts," "reviews," and "orders" sections.
       (2) Initialize every node's weight and score = 0.0 in the hierarchical UIG tree;
Output: The UIG tree of each u with the nodes' inferred interest scores and some new added nodes.
       (1) Extract the set of feature item $I^u\_Ft$ from items in u's profile.
       (2) For each item $i_k$ do
              Calculate the Weight of $i_k$ in $I^u$ : $\omega_{I^u}(i_k)$
              // Traverse the tree from bottom to up.
              For each Node n in the UIG do
                  Calculate $ASS_{i_k}^n$ (Average Semantic Similarity score between n and $i_k$)
                  If $ASS_{i_k}^n$ < $SS_{threshold}$ (set the threshold value of ASS) then
                      Set $ASS_{i_k}^n$ as 0.0
                  Endif
                  Calculate the Weighted Average Semantic Similarity (WSS) score: $WSS_{i_k}^n$ .
```

```
          Record the $k^{th}$ Timed Weight ($TW$) in the set of $n$: $\boldsymbol{TW_{n,k}} < \left\langle WSS_{i_k}^{n}, T_{i_k}^{last} \right\rangle$
        If $n$ is a leaf node then
            // propagate Timed weights up to $n$'s parent node.
              Calculate the $k^{th}$ $TW$ of $n$'s parent: $\boldsymbol{TW_{n\_parent,k} = <T_i^{last}, arg\ max(TW_{n,k}\ and\ sibling\ nodes)>}$
          Else
              $TW_{n\_parent,k} = < T_i^{last}, Average（TW_{n,k}\ and\ sbiling\ nodes）>$
          Endif
      Endfor
      If arg max ( $ASS_{i_k}^{n}$ ) < $SS_{threshold}$ then
      Discovery and add a new node in the $u$'s tree of UIG
      Endif
  Endfor
(3) For each Node $n$ in the tree do
      Calculate Cumulative Weight:  $CW_n = \prod_{i=1}^{|TW_n|}\left(1 - TW_{n,i} \times exp^{-\lambda\left(T^{pre} - T_i^{last}\right)}\right)$
      Mapping node $CW_n$ to a linear value range, 1 to 5.
  Endfor
Return: Output $u$'s UIG with interest scores (1 to 5) assigned for all nodes.
```

4. VALIDATION AND EVALUATION OF THE EXPERIMENT

4.1. Data Acquisition and Preprocessing for Experiments

As mentioned earlier, Dianping.com is a Chinese leading OSN that handles consumers and third-party consumption reviews. It also is one of the biggest platforms in the world. Currently, Dianping has over 200 million users, and over 3 billion items. We chose Dianping.com for experimental validation and evaluation as a result. We collect important items from four representative sections of the user profile “Mine,” and further extracted more fine-grained feature terms from these items.

Due to privacy protection concerns, existing data crawler tools cannot directly obtain user profile data without a user’s authorization. Dianping also does not provide a public API for extracting user profile data. In view of this difficulty, we recruited a number of Dianping’s users to obtain their authorizations, and then extracted online profile data to then construct their UIGs. To avoid the experimental bias caused by the similar interests of friends, the diversity and scope of our Dianping user recruitment is worth considering. The total number of heterogeneous recruited users is 1146, including undergraduates, graduates, on-the-job students of different majors in surrounding universities, and graduates from different industries. We also included teachers and staff in different universities, Dianping users of offline consumer sites, as well as the users recruited online. With the probability sampling method, we randomly selected 50% of the total users. In this way, the authorized users totaled 573 people. Finally, we leveraged a data crawling tool to collect the users' profile data.

In the experiments, to validate the prediction accuracy of this model, we asked them to complete an online questionnaire survey (via www.wjx.cn) to investigate their activity frequency and preferences. The questions were related to the interest topics in the tree. The values of the users' answers in the survey were all in the range 1 to 5 in the Likert scale (5: very interested, 4: moderately interested, 3: a little interested, 2: not very interested, 1: no interest). We used this as ground truth and later compared the data with the inferred scores identified in the subsequent experiments.

Moreover, the collected data were preprocessed according to two requirements: (1) the retained users were required to be active ones with at least 200 items in those sections of their profile, because too sparse a profile could result in an inaccurate prediction in the proposed UIG model. (2) In addition, in some of the questionnaires that were returned, we found that the answers were incomplete or that there were clear contradictions. Such surveys had to be deleted as a result. Thus, 522 of 573 initially gathered users were finally validated and retained. (3) Then, personal information of participating users irrelevant to this research was deleted protect their privacy.

Finally, we analyzed the users' real interest scores on the whole. For the overall nodes on three levels, the global average score was 3.68. In addition, we found that the average interest score decreased with level in monotonic fashion (top level: 4.12, second level: 3.11; third level: 2.74). So on the coarse-grained interest topics of top level, the interests of most users tend to be consistent, while the interests tend to be diverse on the more fine-grained topics of lower levels. And it is reasonable that the average score of the nodes on the top level is relatively higher compared with lower levels.

4.2. Experiments on Scheme Selection and Parameter Adjustments in this Model

For the UIG model and its corresponding algorithm, we first conducted a series of comparative experiments for different scheme choices and parameter adjustments. In these experiments, 100 users were randomly selected for the testing set (about 20% of 522 users).

First, the *mean absolute deviation* (*MAD*) (O'Connor et al. 2002, Seo and Zhang 2001) is employed to select the optimal scheme and parameter values in the experiments. By using *MAD*, the deviation can be calculated between the real and the predicted scores: $\Delta_{u,n}=\left|s_{u,n}^{real}\text{-}s_{u,n}^{pre}\right|$, in which $s_{u,n}^{real}$ denotes the real interest score of a user *u* on an interest node *n*; and $s_{u,n}^{pre}$ denotes the predicted interest score (*CW* has been mapped to the Likert scale). $\Delta_{u,n}$ denotes the absolute Deviation between these two scores. Finally, for all interest scores of all users, the value of $MAD_{U,N}$ is:

$$MAD_{U,N}=\sum_{u=1}^{|U|}\frac{\frac{\sum_{n=1}^{|N|}\Delta_{u,n}}{|N|}}{u} \tag{10}$$

where $|U|$ and $|N|$ represent the total number of users and interest nodes respectively. In this study, the total number of nodes is 167.

Next, by utilizing the $MAD_{U,N}$ metric in Eq. 10, we select the optimal schemes and parameter values for proposed UIG model. The details follow.

4.2.1. The Selection of Item Weight Normalization Schemes

In this experiment, three schemes are designed for item weight normalization for each item i_k, and compared against the $MAD_{U,N}$ values generated from these three schemes.

(a) **Scheme 1**: *Constant item weight***.** The item weight is set with the same constant for each item,. $\omega_{I^u}(i_k) = 1$.

(b) **Scheme 2**: *General item weight normalization*: $\omega_{I^u}(i_k) = tf_{i_k} \Big/ \sum_{k=1}^{|I^u|} tf_{i_k}, i_k \in I^u$.

(c) **Scheme 3**: *Improved weight normalization method proposed earlier.*

After calculations by adopting Schemes 1, 2 and 3, the $MAD_{U,N}$ are 1.79, 1.31 and 1.14 respectively. It is obvious that Scheme 3 is best. So we implemented Scheme 3 in this model.

4.2.2. Determine the Scheme and Parameter Value for Interest Decay Over Time

(a) Scheme 1: Ignoring the interest decay. Take the maximum *TW* value of each node as the final cumulative weight *CW* value. The resultint $MAD_{U,N}$ is 1.68.

(b) Scheme 2: Exponential interest decay. In this scheme, we designed a method with exponential interest decay to calculate the *CW* of node *i*, which decays over time. See Eq. 9. In this experiment, we vary the value of the exponential decay constant λ to obtain the minimum *MAD*. Fig. 6 illustrates the calculated results of $MAD_{U,N}$ for different λ values.

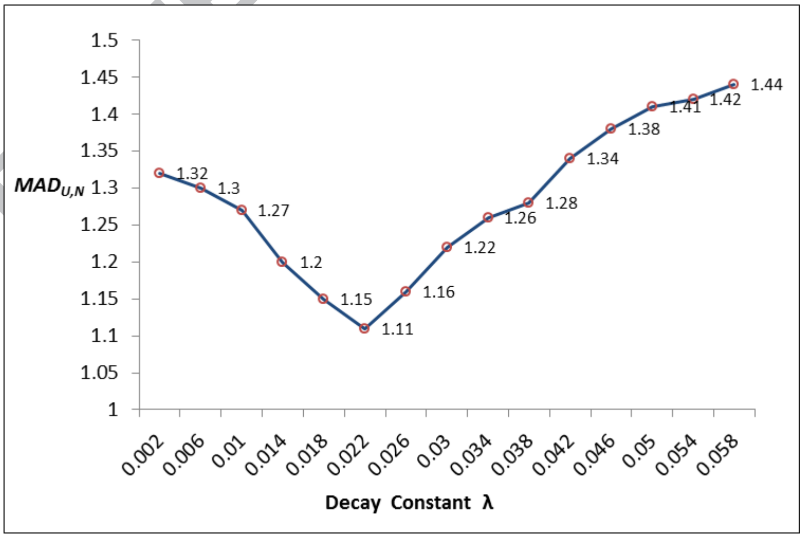


Fig. 6. The calculated results of $MAD_{U,N}$ with different λ values.

After comparison, Scheme 1 (no decay) performs worse than the exponential decay Scheme 2 , in which the maximum $MAD_{U,N}$ is 1.44 and lower than 1.68 obtained from the Scheme 1. Thus, this result validates our intuition that the dynamic nature of user interest over time should be fully considered, and the updated or generated timestamp of each item is an important factor to take into account. As we can see in Fig. 6, λ = 0.022 achieves the lowest $MAD_{U,N}$ 1.11 in the exponential decay scheme. Hence, we use $\lambda = 0.022$ in the following experiments.

4.3. Experiments on Algorithm Performance Compared with Three Benchmarks

In the experiments conducted in this section, three comparative benchmarks are designed based on the proposed UIG model by using different sets of schemes or approaches to verify their impacts on the performance of this model. More details about these benchmarks are described next:

(1) **Benchmark 1**: This benchmark is based on the proposed algorithm. The leaf nodes and non-leaf nodes are not distinguished in the process of applying the bottom-up weight spread in the tree, and the average weight value of the child nodes is spread upward to their parent node.

(2) **Benchmark 2**: This benchmark does not directly utilize an item's name as the only feature to generate the user's interest tree, unlike the fine-grained feature terms, which further extract them from the items. The *ASS* score = 1 if a match is found, else 0 after lexical comparison.

(3) **Benchmark 3**: The algorithm assigns the global average user score 3.68 to each node in the tree uniformly in this benchmark.

Next, to evaluate the performance of the UIG model, we compare the three benchmarks with our algorithm in terms of: prediction accuracy, *MAD*, and the real and predicted relevant interest score.

4.3.1. Accuracy of Interest Prediction

To evaluate interest prediction accuracy, the scores of the real and predicted nodes were divided by using the thresholds for 'Interested' (denoted by *C1*) and 'Not Interested' (denoted by *C2*). Since the scores are all 1 to 5 Likert scale values, we set the classification threshold to 3.0. Thus, $S_{u,n} \geq 3.0$ represents 'Interested' and $S_{u,n} < 3.0$ represents 'Not Interested.' The classical *fusion matrix* method in data mining was employed to measure prediction accuracy, defined as:

$$\frac{t_{c1}+t_{c2}}{t_{c1}+t_{c2}+f_{c1}+f_{c2}} \tag{11}$$

where t_{c1}, t_{c2} represent the number of true positives (*C1*: Interested) and true negatives (*C2*: Not Interested) respectively, and f_{c1}, f_{c2} represent the number of false positives and negatives.

Next, by using the metric shown in Eq. 11, interest prediction accuracy was compared for the proposed UIG model and the three benchmarks. Fig. 7 shows the average prediction accuracy results of all the nodes on each level and all the levels in the tree. We observe that our proposed UIG algorithm performed best for prediction accuracy, compared with the other benchmarks. Since the coarse-grained interest topics on the higher level (Lv1) are more general and tend to be recognized by more users than the fine-grained ones on the lower levels (Lv2 and Lv3). it is reasonable that the average prediction accuracy of Lv1 is comparatively higher than Lv2 and Lv3 in these four models.

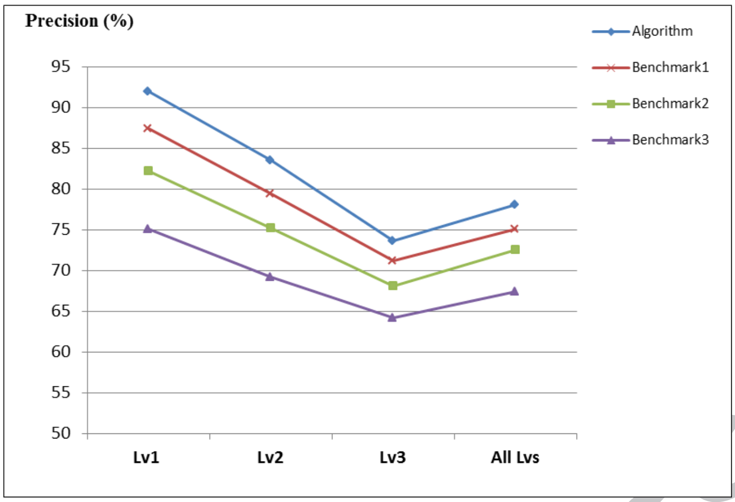


Fig. 7. The average prediction accuracy of all nodes on each level and all levels

4.3.2. Mean Absolute Deviation

In this experiment, the $MAD_{U,N}$ in Eq. 10 is employed to evaluate the performance of prediction deviation for the UIG model and other benchmarks. A lower value for *MAD* indicates superior performance. Similarly, the $MAD_{U,N}$ values of all nodes on each and global levels in the tree of 608 users were calculated for performance comparison. From Fig. 8, the values of $MAD_{U,N}$ are all the lowest on each level, based on our choice to leverage the UIG model and algorithm. Moreover, there is a common trend for the four models: the *MAD* value increased as the level went down. That is because, for the coarse-grained interest topic nodes owned by the higher levels, users were more consistent in their recognition and degree of interest. Naturally, the value calculated by Eq. 10 is lower. Similarly, the interest topics were scattered and more specific on the lower levels, such that users' interests are more inconsistent, and then $MAD_{U,N}$ was naturally comparatively higher.

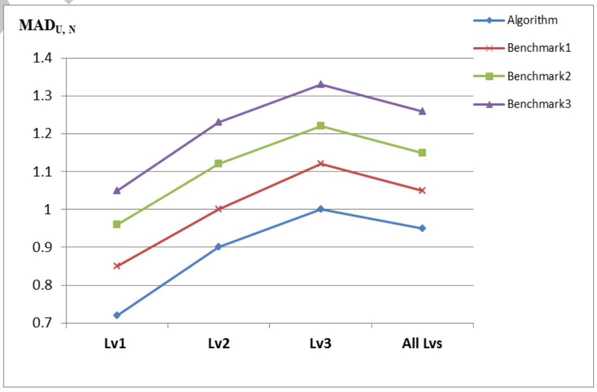


Fig. 8. $MAD_{U,N}$ **of all nodes on each level and on the overall levels**

4.3.3. Pearson Product-Moment Correlation Analysis

The *Pearson product-moment correlation* coefficient (PPMCC) is a widely-used statistical approach to measure the correlation between two continuous variables, and is in the range of [-1, 1]. The larger the absolute PPMCC value is, the stronger the correlation. In this experiment, the PPMCC γ between the real and predicted interest scores on each level for 608 test users was calculated to further validate the prediction accuracy. Fig. 9 presents the results of γ on each level and all levels in the UIG tree by using four benchmark algorithms. It shows that our proposed algorithm achieves the highest γ value at all levels compared with other benchmarks, which expresses the real and predicted interest scores are highly correlated, and also illustrates the algorithm has the distinct advantage on the prediction accuracy. In addition, the values of γ on the Lv1 are all greater than 0.7 from the four algorithms, while the average values of γ are all less than 0.7 on Lv3. This indicates the coarse-grained interest topics can more easily be inferred, and thus the γ value is bigger than those fine-grained topics on lower levels, which also corroborated the experimental results of first accuracy experiment presented previously.

From the results shown in the three experiments, our algorithm outperforms all three other benchmarks. Benchmark 3 is the worst due to its oversimplified method of interest score assignment.

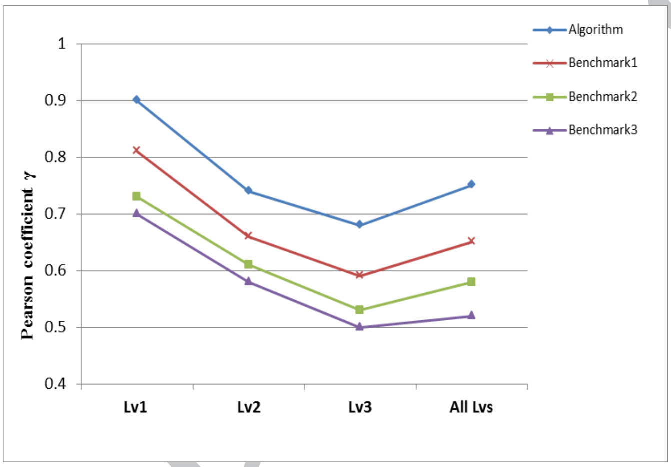


Fig. 9. The results of γ between real and predicted interest scores on each level and overall levels

4.4. Comparison Experiments with Other Hierarchical User Interest Models

4.4.1. Performance Comparisons on Precision, MAD and Pearson Correlation

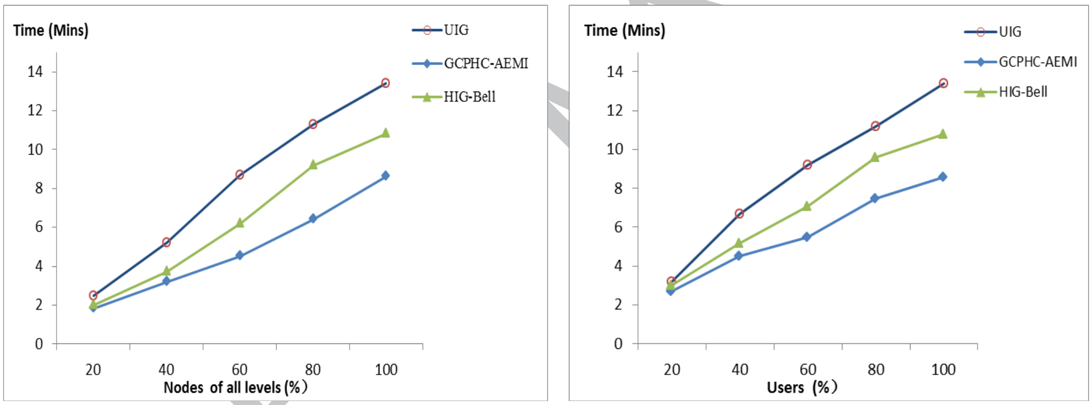
Finally, two typical hierarchical user interest models similar with the proposed UIG were chosen to conduct experiments of performance comparison, GCPHC (Li et al. 2011) and HIG (Kapanipath et al. 2014), respectively. In GCPHC, five correlation functions are used in their algorithms, of which we chose four of the better ones (AEMI, AEMI3, IT, PS) for our performance comparison. Similarly, we adopted two parameters (Bell and Bell log) used in the HIG are model in the comparison experiments. Table 2 shows the results of the performance comparison on three precision metrics, MAD, and Pearson correlation. The values of the precision metrics and Pearson correlation are better when they are larger, and MAD is opposite. Thus, it can be clearly seen that our UIG model outperforms the other two models, GCPHC and HIG, including the various functions and parameters used that are used. In the UIG model, more diverse items are required for the user profile, instead of only web pages or tweets. Also, considering the time decay of interests, the processes of *TW* propagation and *CW* calculation were employed, so this model can infer user interest better. Thus, our results are reasonable based on the above analysis.

Table 2. The result of performance comparison among GCPHC, HIG and the proposed UIG

Models	Precision (%)				MAD				Pearson Coefficient γ			
	Lv1	Lv2	Lv3	All	Lv1	Lv2	Lv3	All	Lv1	Lv2	Lv3	All
GCPHC-AEMI	84.93	79.63	71.21	74.33	0.93	1.08	1.16	1.07	0.81	0.72	0.62	0.66
GCPHC-AEMI3	82.05	76.22	69.48	72.42	0.99	1.14	1.23	1.14	0.77	0.65	0.61	0.62
GCPHC-IT	76.36	69.83	61.27	65.33	1.08	1.16	1.25	1.23	0.72	0.63	0.54	0.57
GCPHC-PS	69.34	58.22	50.58	53.12	1.20	1.26	1.41	1.34	0.62	0.52	0.45	0.50
HIG-Bell	86.11	80.21	72.19	76.32	0.85	0.96	1.09	1.02	0.87	0.75	0.65	0.63
HIG-Bell log	81.70	75.35	67.25	71.31	0.95	1.02	1.14	1.05	0.79	0.69	0.58	0.67
UIG	**91.39**	**83.75**	**75.46**	**80.01**	**0.71**	**0.90**	**1.02**	**0.91**	**0.91**	**0.79**	**0.65**	**0.74**

4.4.2. The Comparison Experiments on Time Consumed

Finally, we evaluated the time consumed by the three algorithms: GCPHC-AEMI (Li et. 2011), HIG-Bell (Kapanipathi et al. 2014) and the proposed UIG. The experimental schemes that we now turn to were designed to include: (1) the number of users was varied with 20% increments in each step; and (2) the nodes on all levels in the tree were increased by 20%. We investigated the time consumed by these three algorithms related to the two schemes separated, as shown in Figs. 10a and 10b. The unit of time is minutes. As we expected, the UIG algorithm took a little more time compared with the others, due to more subtle processes involved in this model. But the time consumed for UIG grew gradually with increases in users and interest nodes, so it has good scalability.



(a) (b)

Fig. 10. Time consumed for three models varied with increase of nodes on (a) all levels and (b) users

From the results of the experiments we conducted, the advantages of our model and algorithm can be summarized. (1) The extracted feature terms are from more diverse items in online user profiles, instead of only the item names, webpages or tweets, so they are very valuable for inferring fine-grained and comprehensive interests. (2) The well-designed TW and CW calculation processes in the proposed model also help to accurately infer implicit interests. Thus, the results of MAD and prediction accuracy were significantly enhanced, especially on the lower levels. (3) Considering the temporally dynamic nature of user interests, the bottom-up weight spread scheme in the tree is also conductive to improving interest prediction accuracy.

5. CONCLUSION

In this study, we proposed a model and a related unsupervised algorithm based on UGC and interaction information extracted from a user's online profile. First, combining the page category of Dianping and Yelp, we constructed a three-level hierarchical UIG tree, covering general interest topics (nodes). Next, we extracted fine-grained features from the items recorded in four representative sections. Then, the semantic similarity was calculated between feature terms and all interest nodes on all levels in the tree of UIG one by one, and then the timed weight of every item was able to be obtained. In addition, the timed weights were propagated from bottom to up in the tree. Taking the characteristic of user's interest decaying over time into consideration, we also designed a scheme of exponential interest decay in this study.

We conducted a series of experiments on scheme selection and parameter adjustment for this model and performance comparisons. In comparison with three benchmarks and two similar hierarchical models on the various metrics including accuracy, MAD and Pearson correlation, our algorithm outperformed other benchmarks and models at all levels of the tree for all users in the collected dataset. In addition, for the metrics on time-consumption and scalability, the proposed algorithm has pretty good performance as well. The experimental results show that the proposed model and algorithm can predict a user's explicit and implicit interests more accurately and comprehensively on both coarse-grained and fine-grained interest topics. Therefore, this research will provide important basic technologies and valuable decision support for social marketing practices, including building accurate user interest profile, personalized ad push-diffusion in OSN, and so on.

Although the proposed UIG model can comparatively accurately infer a user's a wide range of explicit and implicit interests, especially for fine-grained interest topics at the lower levels, how to construct an accurate and efficient model of user interest in OSN is still a challenge. Some problems still need to be further addressed and solved in the future. They include:

Over-reliance on information extracted from individuals. In some cases, just relying on the information extracted from an individual user is not sufficient for accurately inferring his or her interests. Therefore, we are considering to leverage the relational ties between users in OSNs (including unidirectional weak ties or bidirectional strong ties) to further improve the algorithm's accuracy and coverage, especially in the case of inferring user's implicit interests. This idea was inspired by the classic collaborative filtering method, also is based on common sense that friends often share similar interests and preferences. So this approach ought to be useful to address the cold start problem for such a system.

User interests are inferred from positive or neutral feature terms. In the current framework, to simplify the study, user interests are mainly inferred from positive and neutral feature terms. A question worth more deeply exploring is: How do negative terms about "dislike" affect the expression of user interests? This issue is an appropriate direction for future work.

The decay of a user's interest. Another challenge is the decay or change of a user's interest over time. Based on the exponential interest decay approach of proposed in this study, we believe it is worthwhile to design different mechanisms to more accurate reflect the trends that can be discovered in the evolution of a user's interest. This touches on real-world problems related to social marketing.

REFERENCES

Abel, F., Herder, E., Krause, D. 2011. Extraction of professional interests from social web profiles. In L. Ardissono and T. Kuflik (eds.), Proceedings of the Advances in User Modeling Workshop, UMAP 2011, Girona, Spain.

Akcayol, M.A., Utku, A., Aydoğan, E., Mutlu, B. 2018. A weighted multi-attribute-based recommender system using extended user behavior analysis. Electronic Commerce Research and Applications, 28, 86-93.

Asur, S., Huberman, B.A. 2010. Predicting the future with social media. In Proceedings of the ACM International Conferences on Web Intelligence and Intelligent Agent Technology, New York: ACM Press, pp. 492–499.

Banerjee, N., Chakraborty, D., Dasgupta, K., Joshi, A., Mittal, S., Nagar, S., Rai, A., Madan, S. 2009. User interests in social media sites: An exploration with micro-blogs. In Proceedings of the 2009 ACM Conference on Information and Knowledge Management, New York: ACM Press, pp. 1823–1826.

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

Burchell, K., Rettie, R., Patel, K. 2013. Marketing social norms: Social marketing and the social norm approach. Journal of Consumer Behavior, 12, 1, 1-9.

Carmagnola, F., Cena, F., Cortassa, O., Gena, C., Torre, I. 2007. Towards a tag-based user model: How can user model benefit from tags? In Conati, C., McCoy, K., Paliouras, G. (eds.), User Modeling 2007. Lecture Notes in Computer Science 4511, Berlin Heidelberg: Springer, pp. 445–449.

Chan, P. K. 1999. A non-invasive learning approach to building web user profiles. In Proceedings of the 21st ACM SIGKDD International Conference, New York: ACM Press.

Claypool, M., Brown, D., Le, P., Waseda, M. 1997. Inferring user interest. IEEE Internet Computing, 5, 32-39.

Facca, M., Lanzi, L. 2005. Mining interesting knowledge from weblogs: A survey. Data & Knowledge Engineering, 53, 225–241.

Freeman, L.C. 2004. A History of Social Network Analysis, Vancouver, BC, Canada: Empirical Press, 2004.

Garcia Esparza, S., O'Mahony, M.P., Smyth, B. 2013. Catstream: Categorizing tweets for user profiling and stream filtering. In Proceedings of the ACM International Conference on Intelligent User Interfaces, New York: ACM Press, pp. 25-36.

Goel, S., Kumar, R. 2018. Folksonomy-based user profile enrichment using clustering and community recommended tags in multiple levels. Neurocomputing, 315, 425-438.

Han, H., Xu, H., Chen, H. 2018. Social commerce: A systematic review and data synthesis, Electronic Commerce Research and Applications 30 (2018) 38–50.

Hanani, U., Shapira, B., Shoval, P. 2001. Information filtering: Overview of issues, research and system. User Modeling and User-Adapted Interaction, 11, 3, 203–259.

Ho, Q., Yan, R., Raina, R., Xing, E.P. 2012. Understanding the interaction between interests, conversations and friendships in Facebook. arXiv preprint arXiv:1211.0028.

Hogg, T. 2010. Inferring preference correlations from social networks. Electronic Commerce Research and Applications, 9, 29–37.

HowNet. 2019. HowNet's home page. Available at: www.keenage.com.

Kapanipathi, P., Jain, P., Venkataramani, C., Sheth, A. 2014. Hierarchical interest graph from tweets. In Proceedings of the 23rd International Conference on World Wide Web Companion, New York: ACM Press, pp. 1-15.

Karatay, D., Karagoz, P. 2015. User interest modeling in Twitter with named entity recognition. In Proceedings of the 24th International Conference on the World Wide Web, New York: ACM Press, pp. 25-35.

Lazer, D., Pentland, A.S., Adamic, L., Aral, S., Barabasi, A.L., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., Jebara, T., King, G., Macy, M., Roy, D., Van Alstyne, M. 2009. Life in the network: The coming age of computational social science. Science, 323, 5915, 721-723.

Li, F., Li, F. 2001. A new approach measuring semantic similarity. In HowNet 2000, the Journal of Chinese Information Processing, 21, 3, 99-105.

Li, Y., Shiu, Y. 2012. A diffusion mechanism for social advertising over microblogs. Decision Support Systems 54, 2, 9–22.

Li, S., Wu, G., Hu, X. 2011. Hierarchical user interest modeling for Chinese web pages. In Proceedings of the 2011 International Conference on Internet Multimedia Computing and Services, New York: ACM Press, pp. 164-169.

Liang, T.P., Lai, H.J. 2002. Discovering user interests from web browsing behavior: An application to internet news services. In Proceedings of the 35th Annual Hawaii International Conference on Systems Sciences, Los Alamitos, CA: IEEE Computer Society Press, pp. 2718–2727.

Lieberman, H. 1995. Letizia: An agent that assists web browsing. In Proceedings of the 14th International Joint Conference on Artificial Intelligence, 1, San Francisco, Morgan Kauffman, pp. 924-929.

Linden, G., Hanks, S., Lesh, N. 1997. Interactive assessment of user preference models: The automated travel assistant. In A. Jameson, C. Paris, and C. Tasso (eds.), User Modeling, International Centre for Mechanical Sciences (Courses and Lectures), Vienna: Springer, pp. 67-78.

Ma, Y., Zeng, Y., Ren, X., Zhong, N. 2011. User interests modeling based on multi-source personal information fusion and semantic reasoning. In Proceedings of the International Conference on Active Media Technology, Berlin Heidelberg: Springer, pp. 195–205.

Mayer, A. 2009. Online social networks in economics. Decision Support Systems, 47, 3, 169-184.

McCarthy, J F. 2002. Pocket Restaurant Finder: A situated recommender systems for groups. In Proceedings of the Workshop on Mobile Ad-Hoe Communication, 2002 ACM Conference on Human Factors in Computer Systems, Minneapolis, New York: ACM Press.

Mislove, A.E. 2009. Online social networks: Measurement, analysis and applications to distributed information systems. Unpublished thesis, Rice University, Houston, TX.

Nguyen, V., Sriboonchitta, S., Huynh, V. 2017. Using community preference for overcoming sparsity and cold-start problems in collaborative filtering system offering soft ratings. Electronic Commerce Research and Applications, 26, 101-108.

O’Connor M., Cosley, D., Konstan, J.A. Riedl, J. 2002 PolyLens: A recommender system for groups of users. In Proceedings of the Seventh European Conference on Computer Supported Cooperative Work, Berlin Heidelberg: Springer-Verlag, pp. 199-218.

Pazzani, M., Billsus, D. 1997. Learning and revising user profiles: The identification of interesting web sites. Machine Learning, 27, 313-331.

Qiu, F., Cho, J. 2006. Automatic identification of user interest for personalized search. In Proceedings of the 15th International Conference on the World Wide Web. New York, ACM Press, pp. 727-736.

Rajaraman, A., Ullman, J.D. 2011. Mining of Massive Datasets. New York: Cambridge University Press.Raphaeli, O., Goldstein, A., Fink, L. 2017. Analyzing online consumer behavior in mobile and PC devices: A novel web usage mining approach. Electronic Commerce Research and Applications, 26, 1–12.

Sakagami, H., Kamba, T. 1997. Learning personal preferences on online newspaper articles from user behaviors. In Proceedings of the Sixth International Conference on the World Wide Web, New York: ACM Press, pp. 291-300.

Schafer, J.A. Konstan, J. Riedl. 2001. E-commerce recommendation applications. Data Mining and Knowledge Discovery, 5, 1–2, 115–153.

Seo, Y.W., Zhang, B.T. 2001. Learning user’s preferences by analyzing web browsing behaviors.

Artificial Intelligence, 15, 6, 381–387.

Stanford Natural Language Processing Group. 2019. The Stanford Parser: A statistical parser. Stanford University. Available at: nlp.stanford.edu/software/lex-parser.shtml.

Stephen, A.T., Toubia, O. 2010. Deriving value from social commerce networks. Journal of Marketing Research, 47, 2, 1-45.

Su, Q., Chen, L. 2015. A method for discovering clusters of e-commerce interest patterns using click-stream data. Electronic Commerce Research and Applications, 14, 2015, 1–13.

Turban, E., Bolloju, N., Liang, T.P. 2015. Enterprise social networking: Opportunities, adoption, and risk mitigation. Journal of Organizational Computing and Electronic Commerce, 21, 3, 202-220.

Wen, H., Fang, L. Guan, L. 2012. A hybrid approach for personalized recommendation of news on the Web, Expert Systems with Applications 39 (5) (2012), 5806–5814.

White, R.W., Bailey, P., Chen, L.W. 2009. Predicting user interests from contextual information. In Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval, New York: ACM Press, pp. 363-370.

Xu, Z., Lu, R., Xiang, L., Yang, Q. 2010. Discovering user interest on Twitter with a modified author-topic model. In Proceedings of the 2011 IEEE WIC ACM International Conferences on Web Intelligence and Intelligent Agent Technology, Washington, DC: IEEE Computer Society Press, pp. 422–429.

Ying, Q., Chiu, D., Venkatramanan, S., Zhang, S. 2018. User modeling and usage profiling based on temporal posting behavior in OSNs. Online Social Networks and Media, 8, 32-41.

Zabin, J., Brebach, G. 2008. Precision Marketing: The New Rules for Attracting, Retaining, and Leveraging Profitable Customers. New York: John Wiley.

Zheng, J., Wang, S., Li, D., Zhang, B. 2019. Personalized recommendation based on hierarchical interest overlapping community. Information Sciences, 479, 55-75.

Zhu, Z. 2013. Discovering the influential users oriented to viral marketing based on online social networks. Physica A: Statistical Mechanics and Its Applications, 392, 3459-3469.

Zhu, Z., Wang, J., Wang, X., Wan, X. 2016. Exploring factors of user’s peer-influence behavior in social media on purchase intention: Evidence from QQ, Computers in Human Behavior, 63, 980-987.

Zhu, Z., Su, J., Kong, L. 2015. Measuring influence in online social network based on the user-content bipartite graph. Computers in Human Behavior, 52, 184-189.

Conflict of interest statement

The authors declared that we do not have any commercial or associative interest that represents a conflict of interest in connection with the work submitted. In addition, the contents of this manuscript have not been published previously and are not now under consideration for publication elsewhere.

1. Oriented to precision social marketing, a model of user interest graph and its unsupervised algorithm are proposed to accurately and comprehensively infer user’s interests.
2. A three-level hierarchical structure of user interest graph is constructed, which covers a wide range, from coarse-grained to fine-grained interest topics.
3. Fully considering the dynamic nature of user interest over time, a scheme of exponential interest decay is employed.
4. The achievements will provide important basic method and valuable decision supports for precise and personalized social marketing practices.