Multimedia Tools and Applications (2019) 78:32755–32774
https://doi.org/10.1007/s11042-018-7009-y

Clustering based interest prediction in social networks

Xianghan Zheng [1,2] · Wenfei Zheng [1,2] · Yang Yang [1,2] · Wenzhong Guo [1,2] · Victor Chang [3]

Received: 17 April 2018 / Revised: 3 August 2018 / Accepted: 29 November 2018 /
Published online: 8 March 2019

Abstract
Efficient interest prediction for social networks is critical for both users and service providers for behavior analysis and a series of extension services. However, most existing approaches are inefficient, incomplete or isolated. In this paper, we propose combination of Gaussian and Markov approaches (namely, GAM) as typical soft computing technology for interest prediction of social intelligent multimedia systems. GAM model considers "the number of posted messages" as the only parameter, and defines selection logic to implement either Gaussian or Markov based approaches. Our proposed solution takes the advantage of Gaussian model in prediction accuracy and computation complexity, and advantage of Markov model in high availability. Further experiments illustrate that our solution achieves higher prediction accuracy of 94.3% (without considering the influence of swing users), with the best result achieved ever.

Keywords Social network · Clustering · Gaussian mixture model · Multi-Markov chain

1 Introduction

Interest prediction in social network is important for both users (e.g., community participation, activity initiation, etc. [23, 27, 36]), and social network service providers in a series of applications (e.g., behavior analysis, service recommendation, etc. [12, 37]). However, because of 4 V (huge volume, high variety, low value, fast velocity, etc.) characteristics in social multimedia data, feasible and efficient user interest prediction is not a trivial research challenge [18, 29]. On the other hand, similar with people's common life, users in social network can vary from each other in different features. For instance, different users own different number of posted messages (e.g. text, image, video, etc.), online time, and behavior history, etc. Hence,

✉ Xianghan Zheng
xianghan.zheng@fzu.edu.cn

[1] College of Mathematics and Computer Science, Fuzhou University, Fuzhou, China

[2] Fujian Key Laboratory of Network Computing and Intelligent Information Processing, Fuzhou, China

[3] International Business School Suzhou, Xi'an Jiaotong-Liverpool University, Suzhou, China

depending on different feature or environment, social user interest prediction may require different approaches for soft computing.

Existing research is mainly based on three types of information: user registration profile [43], behavior history [3], and social relationship [13]. However, few of them are efficient, complete, and open sourced. This paper considers clustering algorithm as a typical soft computing technology (or computation intelligence) and proposes combination of Gaussian and Markov model (namely, GAM) for social user interest prediction. The clustering technologies are proposed due to the following reasons. First, unsupervised machine learning algorithms are normally computational efficient, especially in big data environments [4, 33]; second, clustering mechanisms take similarity calculation into consideration for better performance enhancement. In particular, we select the combination of Gaussian and Markov models as detailed solution. As described in Section 5, Gaussian content based approach provides accurate results with low computation, whereas Markov status based approach is capable to provide higher availability.

In this paper, the clustering approach proposed in this paper is relevant to soft computing technology. Due to specific implementation scenario for social multimedia data, the clustering prediction of interest requires the participation of computational intelligence. In general, this paper contains three contributions for recent advances in soft computing technology:

- We investigate Gaussian and Markov based clustering approaches (model description, complexity, etc.) respectively for user interest prediction in social networks. Consequently, a compromised GAM model is proposed, which selects either Gaussian or Markov according to the key parameter “Number of posted message”.
- A specific data crawler is developed to collect Sina Weibo as testing dataset. After that, the clustering experiment, strategies selection, and performance evaluations are conducted to show the feasibility and efficiency of proposed solution.
- Through suitable data pre-processing and parameter adjusting, the proposed model achieves 94.3% prediction accuracy. This is the best prediction accuracy achieved ever. Additionally, performance result and model scalability, computation efficiency are discussed to justify our contributions.

Please note that our approach is a generic solution available in other existing social data (Twitter, Facebook, and so on). The paper’s structure is organized as follows. Section 2 investigates social user interest prediction and existing research. Section 3 discusses dataset preprocessing, feature extraction and user annotation. Section 4 investigates GMM and MCM approaches respectively and introduce our proposed GAM model. Section 5 illustrates experiment, analysis and discussion. Finally, Section 6 summaries the paper.

2 Social interest prediction

2.1 Social network and user interest

Online social networks have become major platforms for internet users to post multimedia messages (e.g. text, picture, video, etc.), discuss and share interesting topics [33]. In social network, interest is usually represented by posted messages describing the event or willingness such as what want to do or buy, where want to

go, or who want to meet, follow or vote for [10]. Therefore, interest exploration in social networks is an important part of user behavior analysis, sine it can provide support for a series of extension services such as community detection targeted advertisement, personalizing recommendation and so on [34].

This paper investigates and collects dataset from Sina Weibo, over 350 million users and the eighth most frequently visited social network in the world until Dec 2017 [25, 31]. Upon this dataset, the investigation and further experiment is convincible and scalable. Therefore, this is a very relevant paper for the soft computing research on social network and multimedia big data.

2.2 Existing works

In industry, both Twitter and Facebook initiate their research project implementing machine learning technologies for user behavior analysis, according to their annual report [32]. However, the details are unknown to the public.

In academia, the initial attempt is to explore relevant messages entered by user as interest information so as to establish user interest prediction model [1] [6]. Abel F et al. [2] extend users' basic information through tagged user profiles, and develop a cross-system user model to find user interest and improve recommendation quality. Xu et al. [35] filters interest-unrelated noisy posts according to aggregated user profiles, and to some extents, discovers user interest. These approaches are based on user's registration information; however, the result may not be accurate due to incomplete user information entered.

Besides, there are some other method based on social relationship analysis. Xiaoling S et al. [28] propose an agent-based interest awareness model that considers social ties formed or reinforced between two individuals if they have similar interest. Xiao H et al. [15] capture various social features and investigate social inference based on interest similarity to realize users interest prediction. Saber Shokat Fadaee et al. [11] convert social network into Bernoulli based unweighted structure model, and predict user interest category according to structural difference between different categories of networks. Norietal et al. [21] import graph theory to model user time-evolving behavior, and predict user interest category via similarity computation. However, these approaches are incomplete because social relationship is only a part of feature for interest exploration.

Some other approaches are based on feature exploration on social network content. For example, Attenberg et al. [5] predict user interest through analyzing message content posted. Banerjee et al. [7] collect Twitter data and apply statistical and mining techniques to explore user interest distribution on categories, e.g., food, sport, movie, etc. Literature [19] considers the imbalanced data of social users and introduces an weighted ELM based on the overall distribution (ODW-ELM) model for predicting users future interests. [40] considers the evolution of user interests and utilizes semantic information from knowledge bases such as Wikipedia to predict user future interests and overcome the cold item problem. [24] proposes a multilevel deep belief network learning-based model for users consumption preferences, based on interaction between the preferential behaviors of users. Our previous research [42] also proposes a Markov chain model on clustered users to predict user interest. However, the disadvantage of above approaches is that most of them define complicated

computation logic that cause a lot of system burden. Besides, these approaches only classify each user into a specific interest category, while in reality each user may have multiple interest. Additionally, none of these approaches achieve excellent performance result (most are between 60% - 80% in prediction accuracy).

Generally, this paper is the extension of our previous work [42] with a few significant improvements:

1. This paper integrates Gaussian and Markov based approaches, which achieves lower computation complexity and better performance outputs.
2. Both theoretical and experiment illustrate that via only inspection of the parameter "number of posted message", our proposed solution is capable to select optimized handling logic. This makes the model implementation easy.
3. 94.3% prediction accuracy could be obtained with suitable parameter adjusting (weed out the influence by swing users). This is the best result ever.

3 Dataset analysis

3.1 Dataset collection for social networks

Similar as most social media platforms, the public Weibo developer API (specifically, user_timeline API) only provides the downloadling functionality on the recent messages of authorized users. This is considered as an obstacle to the process of data collection. To solve this problem, specific data crawler and feature collection mechanism are developed. Specifically, we manually select 20 interest categories source data that contains 100 normal users (who post, repost, or comment frequently) as data source. After that, a specific data crawler is developed for dataset collection. The data crawler contains two classes: WeiboCrawler for collecting user related information, expecially posted messages, followee's ID, etc.; and FolloweeCrawler class that collecting followee's posted messages. Finally, 30,116 Weibo users with around 17 million messages are acquired (from 20th, Jan, 2017 to 1st, April, 2017) are extracted.

Figure 1 illustrates the distribution of "the number of posted messages". It shows that most normal users post/repost 250-550 messages (including text, image, video, etc.) in around 70 days.

3.2 Feature vector extraction

After dataset collection, feature vector could be generated according to following steps [22]:

1. Word Segment and Frequency Statistics. Via filtering image and video content and deploying the Chinese Institute of Computing Segmentation System (ICTCLAS) [8, 30], it is capable to extract separated words from Weibo message. After that, according to affiliated TF-IDF (term frequency–inverse document frequency) API [9], the top 50 keywords for each 20 predefined interest category could be obtained. Consequently, the total number of keywords is $20 * 50 = 1000$.
2. De-duplication and Feature Vector Generation. After manual re-inspection to reduce redundancy, we achieve 579 keywords, based on which feature vector could be generated with dimension of 1*579.

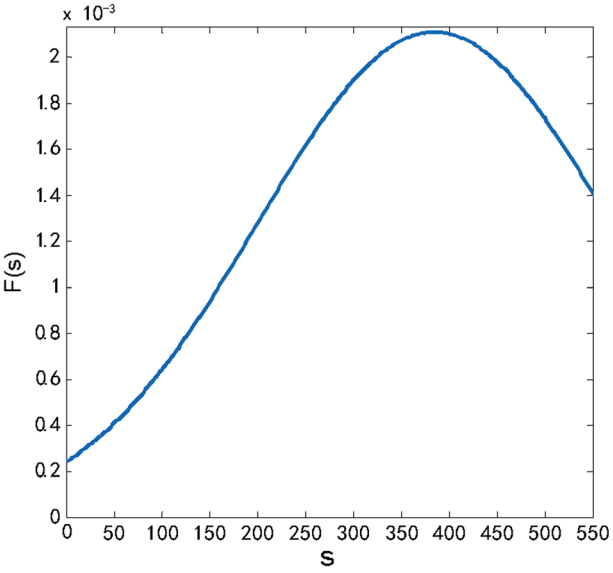


Fig. 1 The distribution of "the number of posted messages"

3.3 User annotations

Among 30,116 users, we randomly select 4000 users and assign three volunteers to handle the annotation work, marking user interest category according to the message history. The marking behavior of three volunteers is not interfered each other. In case one user is marked in different categories, the majority voting is implemented for suitable decision. Finally, user number and corresponding category is illustrated in Table 1.

4 Solution

Figure 2 illustrates the overview of proposed solution. After feature vector generation (described in Section 3), clustering algorithms (e.g., Markov chain model, GMM model and so on) are applied to construct prediction model.

Table 1 User number and corresponding Category

Category	Number	Category	Number
Entertainment	721	Health	256
Finance	68	Cartoon	79
Sports	74	Movie	248
Culture	524	Travel	31
Fashion	182	Food	52
Constellation	96	Pets	55
Tip-off	341	Pictures	86
Joke	63	Music	114
Emotion	174	Hallyu	53
Technology	346	Embarrassment	272

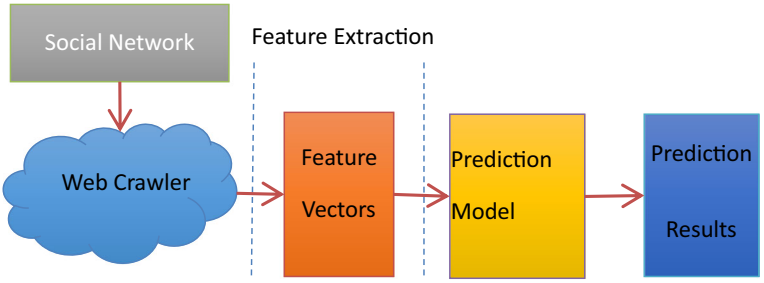


Fig. 2 System overview

4.1 GMM based prediction

4.1.1 Gaussian mixture model

According to [38], Gaussian mixture model is described as the following fomula:

$$p(x) = \sum_{k=1}^{K} \pi_k N(x|\mu_k, \sum k) \tag{1}$$

Where $N(x|\,\mu_k, \sum k)$ is density function, μ_k, $\sum_k$and$\pi_k$are corresponding mean, covariance and mixing coefficient respectively. According to sum and product rule, the marginal density is:

$$p(x) = \sum_{k=1}^{K} p(k)p(x|k) \tag{2}$$

Supposed that the total number of messages user published is s, and$s \sim N(\mu, \sigma)$; the classification number k and s are independent each other, here is the Theorem:

> **Theorem 1**: the prediction accuracy$p(x)$ is a monotone increasing function with the increasing number of s.
> **Proof**: $p(x|\,k)$ in formula (2) can be transformed to$p(x|\,k, s)$, as follows:

$$p(x|k,s) = \frac{p(x,k,s)}{p(k,s)} \tag{3}$$

Since the parameter $k$and $s$are independent each other, $p(k,s) = p(k) \times p(s)$,$p(s|\,x,k) = p(s|\,x)$, formula (3) can be transformed to:

$$\begin{aligned} p(x|k,s) &= \frac{p(x,k) \times p(s|x)}{p(k) \times p(s)} \\ &= \frac{p(x,k) \times \frac{p(s,x)}{p(x)}}{p(k) \times p(s)} = \frac{p(x,k)}{p(x) \times p(k)} \times p(x|s) \end{aligned} \tag{4}$$

Where $\frac{p(x,k)}{p(x) \times p(k)}$ is not affected by s and $p(x|\,s)$ is increased with the increasing number of s, therefore the theorem is proved.

4.1.2 EM steps

The maximum likelihood of Formula (1) is illustrated in the following formula:

$$\ln p(X|\pi,\mu,\Sigma)=\sum_{n=1}^{N}\ln\left\{\sum_{k=1}^{K}\pi_k N(x_n|\mu_k,\Sigma_k)\right\} \tag{5}$$

where $X=\{x_1,...,x_N\}$.

Additionally, EM algorithm [20, 39] is implemented with the following steps:

1. Initialize μ_k, $\sum_k$and π_k, and calculate initial likelihood.
2. E-step:

$$\gamma(z_{nk})=\frac{\pi_k N(x_n|\mu_k,\Sigma_k)}{\sum_{j=1}^{K}\pi_j N\left(x_n|\mu_j,\Sigma_j\right)} \tag{6}$$

3. M-step:

$$\mu_k^{new}=\frac{1}{N_k}\sum_{n=1}^{N}\gamma(z_{nk})x_n, \tag{7}$$

$$\Sigma_k^{new}=\frac{1}{N_k}\sum_{n=1}^{N}\gamma(z_{nk})\left(x_n-\mu_k^{new}\right)\left(x_n-\mu_k^{new}\right)^T \tag{8}$$

$$\frac{\pi_k N(x|\mu_k,\Sigma_k)}{\sum_{j=1}^{K}\pi_j N\left(x|\mu_j,\Sigma_j\right)}, \tag{9}$$

Where

$$N_k=\sum_{n=1}^{N}\gamma(z_{nk}). \tag{10}$$

4. Log likelihood Evaluation

$$\ln p(X|\pi,\mu,\Sigma)=\sum_{n=1}^{N}\ln\left\{\sum_{k=1}^{K}\pi_k N(x_n|\mu_k,\Sigma_k)\right\} \tag{11}$$

E-Step 2 would be returned until convergence criterion is satisfied. Consequently, optimized parameter with result value can be obtained.

4.1.3 Time complexity of GMM approach

Theorem 2: the time complexity of GMM algorithm for social interest prediction is $0(n^2k)$, given interest category number k and user posted message number n.

Proof: For initialization the variables of k initial categories, the execution time is $0(k)$; for E-step calculation, the execution time is $0(nk)$; for M-step calculation, the execution time is $0(n^2k)$; for maximum likelihood function, the execution time is $0(nk)$. Therefore, GMM time complexity is $0(n^2k)$. The theorem is proved.

4.1.4 Computation complexity

Theoretically, with the increasing number of s, the value of $p(x)$ increases (refer to the Matlab simulation result in Fig. 3). It is obviously that (1) GMM is capable to achieve high prediction result (for example, it reaches over 0.9 when user posted messages is more than 375); (2) however, GMM may not work efficiently in case that user posted messages is not enough (for instance, the prediction accuracy would be less than 0.7 when s is less than 175). Therefore, in order to further improve prediction accuracy, it might be necessary to introduce some other methods.

4.2 Markov chain model (MCM)

GMM based interest prediction is content based approach that require as much as user posted message as possible. This might be not efficient for users when posted message is inadequate. On the other hand, Markov model is status based prediction approach that might generate reliable result as long as its status chain has been constructed [17]. Therefore, Markov based interest prediction might be implementable for further improvement of prediction accuracy.

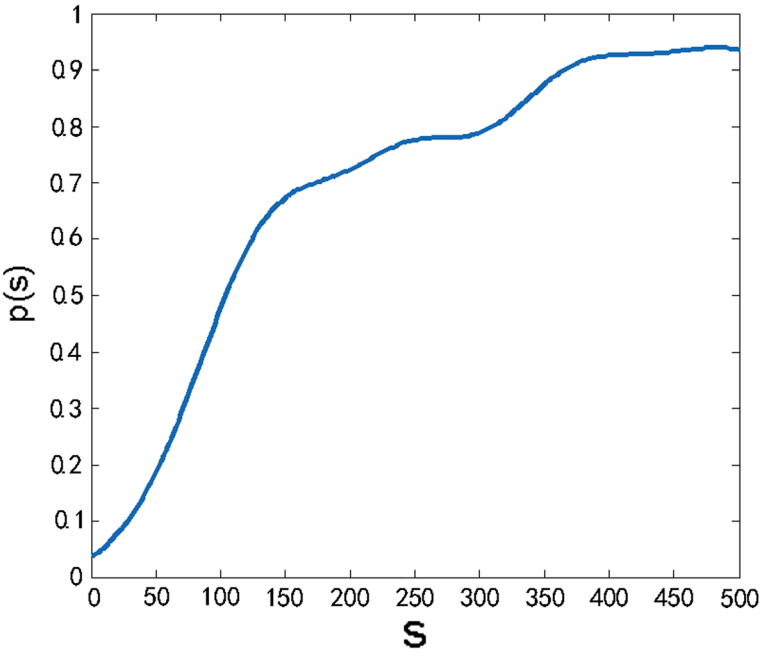


Fig. 3 Simulation result of effect of s to p(x)

4.2.1 Markov chain model

Our previous work [42] has modeled user interest prediction in social network as a triplet $MC = <X, A, \lambda>$, in which A is transition rate matrix:

$$A = \left(p_{ij}\right) = \begin{bmatrix} P_{11} & P_{12} & \dots & P_{1j} & \dots & P_{1n} \\ P_{21} & P_{22} & \dots & P_{2j} & \dots & P_{2n} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ P_{i1} & P_{i2} & \dots & P_{ij} & \dots & P_{in} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ P_{n1} & P_{n2} & \dots & P_{nj} & \dots & P_{nn} \end{bmatrix} \tag{12}$$

Where $p_{ij} = P(X_t = x_j | X_{t-1} = x_i)$ means transition probability from x_i to x_j; λ refers to initial state distribution:

$$\lambda = (p_i) = (p_1, p_2, ..., p_n) \tag{13}$$

After that, via maximum likelihood calculation, each parameter in Markov model is capable to be estimated:

$$p_{ij} = \frac{S_{ij}}{\sum_{j=1}^{n} S_{ij}} \tag{14}$$

$$p_i = \frac{\sum_{j=1}^{n} S_{ij}}{\sum_{i=1}^{n}\sum_{j=1}^{n} S_{ij}} \tag{15}$$

However, our previous method only classify each user into a specific interest category. For further multiple interest prediction, this paper defines multi-Markov chain model and its corresponding solution.

> Definition 1: The multi-Markov chain (m-MCM) based user interest model is represented as a quaternion: $<X, K, P(C), MC>$, where X is discrete random variable in range $\{x_1, x_2, ..., x_n\}$, each xi refers to interest eigenvalue, $C = \{c_1, c_2, ..., c_k\}$ refers to a group of user interest categories with the number of k, $P(C = c_k)$ refers to probability of i-th category, $MC = \{mc_1, mc_2, ..., mc_k\}$ represents multiple Markov chains and each element mc_k belongs to category c_k. Therefore, the transition matrix A_k could be represented:

$$A_k = (p_{kij}) = \begin{bmatrix} P_{k11} & P_{k12} & \dots & P_{k1j} & \dots & P_{k1n} \\ P_{k21} & P_{k22} & \dots & P_{k2j} & \dots & P_{k2n} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ P_{ki1} & P_{ki2} & \dots & P_{kij} & \dots & P_{kin} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ P_{kn1} & P_{kn2} & \dots & P_{knj} & \dots & P_{knn} \end{bmatrix} \tag{16}$$

The initial state distribution, λ_k is represented as follows:

$$\lambda_k = (p_{ki}) = (p_{k1}, p_{k2}, ..., p_{kn}) \tag{17}$$

$$p_{kij} = \frac{S_{kij} + \alpha_{kij}}{\sum_{j=1}^{n} (S_{kij} + \alpha_{kij})} \tag{18}$$

$$p_{ki} = \frac{\sum_{j=1}^{n} (S_{kij} + \alpha_{kij})}{\sum_{i=1}^{n} \sum_{j=1}^{n} (S_{kij} + \alpha_{kij})} \tag{19}$$

Where k and S_{kij} refer to number of interest categories and status pair respectively; a_{kij} represents background knowledge in Bayesian estimation [16, 41].

After that, we calculate the similarity δ_{kl}among each two users' transfer matrixes, with the calculation formulas:

In case the value of δ_{kl} is large or infinite, two users are regarded in one interest category. The merging formulas could be illustrated:

$$\delta_{kl} = \frac{S\text{imilarity}(mc_k, mc_l)}{S\text{imilarity}(mc_k, mc_l) + S\text{imilarity}(mc_l, mc_k)} = 2 \tag{20}$$

$$p_{(k+l)ij} = \frac{S_{kij} + S_{lij} + \alpha_{(k+l)ij}}{\sum_{j=1}^{n} (S_{kij} + S_{lij} + \alpha_{(k+l)ij})} \tag{21}$$

$$p_{(k+l)i} = \frac{\sum_{j=1}^{n} (S_{kij} + S_{lij} + \alpha_{(k+l)ij})}{\sum_{i=1}^{n} \sum_{j=1}^{n} (S_{kij} + S_{lij} + \alpha_{(k+l)ij})} \tag{22}$$

Consequently, multi-Markov chain for user interest prediction could be constructed.

4.2.2 Computation complexity

Theorem 3: the time complexity of MCM based approach is $O(m^3n^2)$, given m as user interest enginvalue and n as the total number of user messages.

Proof: The MCM algorithm contains two part: the initial part and the circulation part. In initiation part that calculates p_{ij}andp_i, and transforms user data into Markov Chain, the execution time is $O(mn^2)$.

In circulation part, the maximum cycle time is m because there are always two Markov chains merged (or exit the loop, the algorithm ends) for every cycle. Additionally, the calculation of

similarity degree between different pairs listed in descending sequence costs $O(m^2n^2)$ (If ignore the sorting operation time). In the calculation that merges the two Markov chain with the maximum similarity degree, the execution time is $O(m^2n^2)$. Then the execution time of the circulation part is $O(m*m^2n^2)$, that is, $O(m^3n^2)$. In combination, the time complexity of MCM based approach is $O(m^3n^2)$. And the theorem is proved.

4.3 Gaussian and Markov approach (GAP)

From Section 4.1 and 4.2,it obviously that both GMM and Markov based approaches have their own advantages. GMM model is content based approach that provides lower computation complexity, while Markov model is status based approach that may not require a large amount of user message as long as the status matrix could be constructed and stablized.

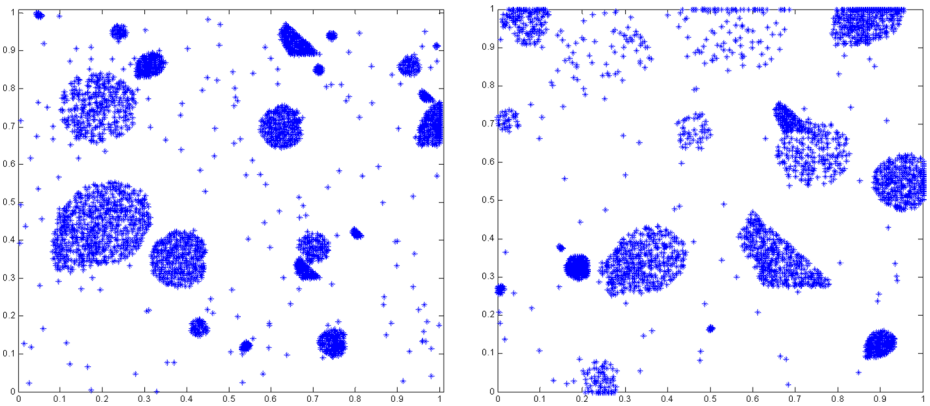
Therefore, one feasible combination can be: (1) set a predefined number w; (2) when the number of user's posted messages $s > w$, implement GMM based prediction; while $s < w$, implement MCM based prediction; (3) adjust the value of w until the best prediction accuracy achieved.

5 Experiment and evaluation

The experiment and evaluation work are described from four aspects. First, the clustering result between GMM and MCM is described; after that, the integration strategy is investigated and tested; additionally, the performance evaluation is conducted with a few existing algorithms; and finally, we discuss the model implementation and scalability. Based on the collected data set, 4000 out of 30,116 users are randomly selected as experiment users. The experiment is conducted in Matlab [14] environment.

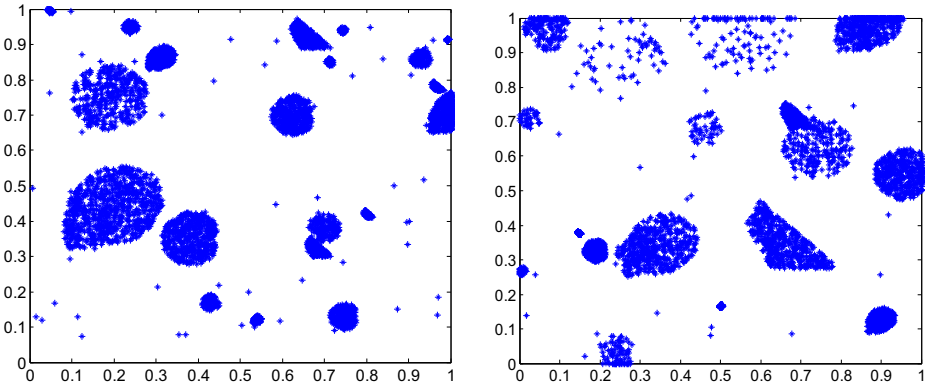
5.1 Clustering result

Figure 4(a) and (b) show clustering results of GMM and MCM respectively. This result contains noise which is the swing users (whose interest categories are difficult to be



(a) GMM model (b) MCM model

Fig. 4 The Clustering Result



(a) GMM after Denoised (b) MCM after Denoised

Fig. 5 The Clustering Results after Denoised

determined from). After noise filter (component analysis method provided in MATLAB), we obtain clearer clustering results illustrated in Fig. 5, which contains 3835 valid users.

It is obvious that user classification result in Fig. 5(a) (boundary among clusters is better splitted) is better than that in Fig. 5(b) (certain clusters are scattered and difficult to be determined). Specifically, the user can be accurately divided into 20 categories in Fig. 5(a), whereas only 14 categories can be distinguished in Fig. 5(b). Therefore, the cluster graph shows that GMM approach has better capability in terms of splitting the boundary of each interest category compared with MCM approach.

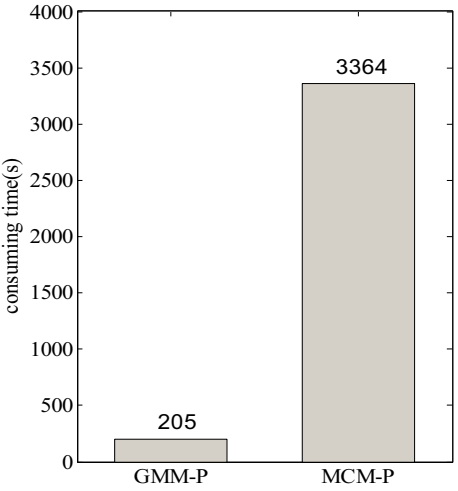
More detailed, the clustering result in 20 categories is listed in Table 2. Although some classification might be wrong (for instance, some users classified as 'Entertainment' categories may not belong to this category,etc.), the high gap between GMM and MCM indicates the advantage of adopting GMM for clustering analysis in crowd intelligence.

On the other hand, the time consumption between GMM and MCM is also compared. As shown in Fig. 6, MCM algorithm takes almost 16 times longer than GMM.

Table 2 The Clustering Results

Category	Manually Classified Number	GMM	MCM	Category	Manually Classified Number	GMM	MCM
Entertainment	721	699	612	Health	256	245	213
Finance	68	67	86	Cartoon	79	76	63
Sports	74	72	62	Movie	248	235	253
Culture	524	509	503	Travel	31	24	43
Fashion	182	175	168	Food	52	51	44
Constellation	96	90	80	Pets	55	50	43
Tip-off	341	319	353	Pictures	86	80	91
Joke	63	61	84	Music	114	111	125
Emotion	174	163	196	Hallyu	53	49	44
Technology	346	330	367	Embarrassment	272	265	291

Fig. 6 Time Consumption



5.2 Combination strategy

For exploring optimized integration, we further investigate clustering error in both GMM and MCM methods, and find the obvious feature difference in the number of posted message (including text, image, video, etc.) between these two approaches.

Table 3 Error Users Analysis in GMM and MCM Respectively

Category	GMM based Prediction			MCM based Prediction		
	Error User Number Using GMM	Average number of posted message in Error users	Error User in GMM while correct in MCM	Error User Number in MCM	Error User in MCM while correct In GMM	Average number of posted message
Entertainment	52	118	35	175	127	269
Finance	6	52	6	14	8	356
Sports	5	134	5	27	23	431
Culture	6	156	4	115	104	465
Fashion	35	36	17	50	47	226
Constellation	11	68	6	18	14	389
Tip-off	46	86	29	79	78	364
Joke	6	169	0	23	17	405
Emotion	23	96	17	22	18	359
Technology	29	139	23	32	29	337
Health	35	208	29	33	30	468
Cartoon	12	66	6	18	12	481
Movie	29	68	23	19	14	427
Travel	6	34	0	9	4	431
Food	6	89	6	14	9	296
Pets	5	130	5	10	6	376
Pictures	12	106	6	18	11	419
Music	16	66	12	37	26	437
Hallyu	6	55	6	11	7	416
Embarrassment	35	135	29	38	29	399

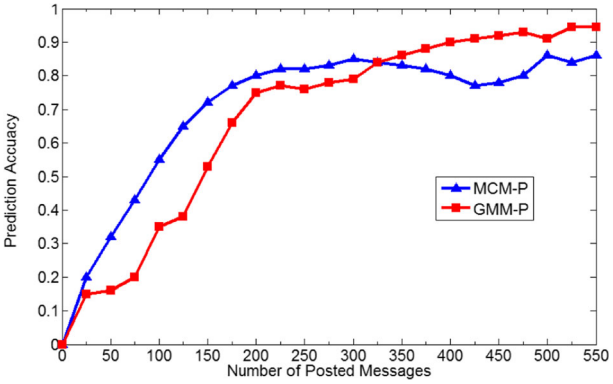


Fig. 7 Effect of different number of posted messages

5.2.1 Clustering error analysis

According to previous clustering result, the error data can be listed, as shown in Table 3. We check each error user respectively, and find that most error users produced by GMM contain less posted message (between 34 and 208 in each category) than error users generated by MCM. Therefore, the number of posted message may probably a distinguished feature difference between GMM and MCM approaches.

5.2.2 The number of post message effect

Furthermore, this section investigates the features of user posted messages (with the total number between 0 and 50, 50–100, 100–150, 150–200, 200–250, 250–300, 300–350, 350–400, 400–450, 450–500, 500–550, and 550 plus) and discusses the effect of "The Number of Post Message" in the prediction accuracy. We randomly select 20 users in each interval and test the prediction accuracy with GMM and MCM approaches respectively. The results in Fig. 7 shows that (1) GMM algorithm is more accurate than MCM when the number of post

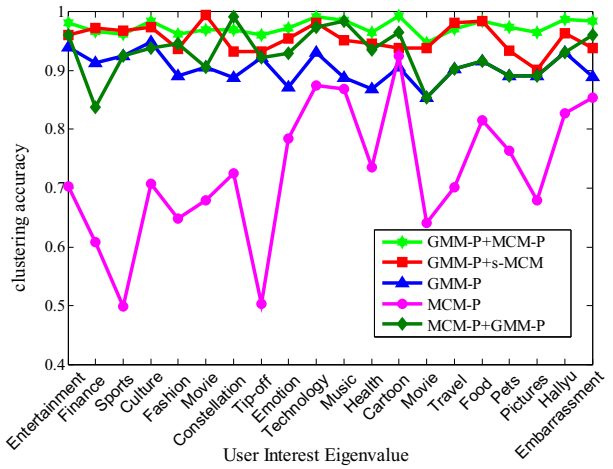


Fig. 8 Performance of Evaluation of Different models

Table 4 Confusion matrix evaluation

		Predicted	
		Positive	Negative
Actual	Positive	93.49%	6.51%
	Negative	5.18%	94.82%

messages is larger than 300; (2) MCM approach is still capable to achieve much better performance in case that the number of post messages is between 50 and 100 and 100-150.

Therefore, we set the threshold value in this paper as 150, and select to implement GMM approach when above threshold, and implement MCM when below this value. The prediction accuracy is further improved to around 95% in each of 20 categories (See Fig. 8).

5.3 GAM performance

Furthermore, we compare MCM solution with GAM, GMM based solution, and also and traditional solutions such as LIBSVM, K-Means algorithms provided by Weka tool [26]. Table 4 shows that proposed solution is 93.49% positive and 94.82% negative classified. In other words, our prediction can reach 93.49% for true positive and 94.82% true negatives. Table 5 further calculate precision, recall, and F-measure values, which are always above 0.9. Finally, the comparison between SVM and other classifiers illustrates that our GAM solution can achieve the highest prediction accuracy (shown in Table 6).

5.4 Discussion

Three topics of discussions to justify our research contributions in recent advances in crowd intelligence are as follow:

(2) Model Feasibility and Scalability: the compromised GAM model (integrates Gaussian and Markov based clustering approaches) is theoritically feasible and proved/validated via a series of experiment. Additionally, the proposed solution, with few revision, is scalable for any other social networks (e.g. Facebook, Titter, etc.).

(3) Computation Efficiency: As compromise of GMM and MCM, the computation efficiency of GAM depends on the ratio of users with "the number of posted message" below or above a predefined threshold value. Since most normal users in social network do post messages more than predefined threshold (As seen in Fig. 1), our proposed GAM solution would cause only a little bit more computation burden than GMM approach. This is regarded to be acceptable.

Table 5 Classification evaluation

	Precision	Recall	F-measure
Positive	0.9432	0.9349	0.9393
Negative	0.9364	0.9482	0.9433

Table 6 Comparison with traditional classifiers

Classifier	Precision		Recall		F-measure	
	Positive	Negative	Positive	Negative	Positive	Negative
GAM	0.943	0.936	0.935	0.948	0.939	0.943
GMM	0.923	0.907	0.894	0.925	0.918	0.921
MCM	0.883	0.857	0.865	0.885	0.874	0.866
LibSVM	0.840	0.855	0.859	0.836	0.849	0.846
K-means	0.845	0.834	0.831	0.848	0.838	0.841

(4) Performance Enhancement: Due to the different dataset and environment setting, it is difficult to directly compare the performance with existing works. However, it is obvious that our solution achieves the highest result due to two reasons: firstly, existing work take either tag / limited content, or only social relationship into consideration, while our solution considers all posted messages; secondly, our proposed solution considers "the number of posted message" as the only key feature, and is capable to select optimized handling mechanism. In summary, we have greatly improve the prediction accuracy from 60%-80% (See reference [5, 11, 15, 21, 28, 35]) to 94.3% in our work.

6 Conclusions

User interest prediction in social network has become hot topic in both academia and industry. This paper introduces clustering approaches to achieve soft computing (or computational intelligence), specifically GMM, MCM and finally proposes a GAM solution to predict user interest in social networks. We have conducted a series of experiments and analysis to show that our proposed GAM solution is feasible, efficient, and achieving a higher prediction accuracy. Comparing with other algorithms or existing work, our proposed solution also contains acceptable computation complexity. We demonstrate our work for recent advances in soft computing and justify our research contributions by applying different methods to meet prediction challenges for social intelligent multimedia systems.

References

1. Abel F, Arajo S, Gao Q et al (2011) Analyzing cross-system user modeling on the social web.[J]. Lect Notes Comput Sci 6757(2-3):28–43
2. Abel F, Herder E, Houben GJ et al (2013) Cross-system user modeling and personalization on the social web[J]. User Model User-Adap Inter 23(2-3):169–209
3. Agarwal V (2013) Bharadwaj K K. a collaborative filtering framework for friends recommendation in social networks based on interaction intensity and adaptive user similarity[J]. Soc Netw Anal Min 3(3):359–379
4. Anderberg MR (2014) Cluster Analysis for Applications: Probability and Mathematical Statistics: A Series of Monographs and Textbooks[M]. Academic press
5. Attenberg J, Pandey S, Suel T (2009) Modeling and predicting user behavior in sponsored search[C]// Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM: 1067–1076
6. Baltrunas L, Ricci F (2014) Experimental evaluation of context-dependent collaborative filtering using item splitting[J]. User Model User-Adap Inter 24(1-2):7–34

7. Banerjee N, Chakraborty D, Dasgupta K et al. (2009) User interests in social media sites: an exploration with micro-blogs[C]//Proceedings of the 18th ACM conference on Information and knowledge management. ACM 1823–1826
8. Carpineto C (2012) Romano G. a survey of automatic query expansion in information retrieval[J]. ACM Comput Surv (CSUR) 44(1):1
9. Erra U, Senatore S, Minnella F et al (2015) Approximate TF–IDF based on topic extraction from massive message stream using the GPU[J]. Inf Sci 292:143–161
10. Facebook, in: http://www.facebook.com/
11. Fadaee SS, Farajtabar M, Sundaram R et al (2015) On the network you keep: analyzing persons of interest using Cliqster[J]. Soc Netw Anal Min 5(1):1–14
12. Felix W, Zhengming L, Mung C et al (2016) On the efficiency of social recommender networks[J]. IEEE/ACM Trans Networking 24(4):2512–2524
13. Gonzlez E, Turmo J (2015) Unsupervised ensemble minority clustering[J]. Mach Learn 98(1-2):217–268
14. Grewal MS, Andrews AP (2014) Kalman filtering: Theory and Practice with MATLAB[M]. Wiley
15. Han X, Wang L, Crespi N et al (2015) Alike people, alike interests? Inferring interest similarity in online social networks[J]. Decis Support Syst 69:92–106
16. Heckerman D, Geiger D, Chickering DM (1995) Learning Bayesian networks: the combination of knowledge and statistical data[J]. Mach Learn 20(3):197–243
17. Herrmann JW (2015) Predicting the performance of a design team using a Markov chain model[J]. Eng Manag, IEEE Trans 62(4):507–516
18. Kunaver M, Porl T (2017) Diversity in recommender systems A survey[M]. Elsevier
19. Luo X, Jiang C, Wang W, et al. (2018) User behavior prediction in social networks using weighted extreme learning machine with distribution optimization ☆[J]. Futur Gen Comput Syst
20. Melnykov V, Melnykov I (2012) Initializing the EM algorithm in Gaussian mixture models with an unknown number of components[J]. Comput Stat Data Anal 56(6):1381–1395
21. Nori N, Bollegala D, Ishizuka M (2011) Interest prediction on multinomial, time-evolving social graph[C]// IJCAI 11: 2507–2512
22. Phan XH, Nguyen CT, Le DT et al (2011) A hidden topic-based framework toward building applications with short web documents[J]. Knowl Data Eng, IEEE Trans 23(7):961–976
23. Scott J (2012) Social network analysis[M]. Sage
24. Sharma P, Rathore S, Park JH (2017) Multilevel learning based modeling for link prediction and users' consumption preference in Online Social Networks[J]. Futur Gen Comput Syst
25. Sina, in: http://www.sina.com.cn/
26. Singhal S, Jena MA (2013) Study on WEKA Tool for Data Preprocessing, Classification and Clustering[J]. Int J Innov Technol Exp Eng 2(6)
27. Statista, in: http://www.statista.com/
28. Sun X, Lin H, Xu K (2015) A social network model driven by events and interests[J]. Expert Syst Appl 42(9):4229–4238
29. Tang J, Liu H (2014) An unsupervised feature selection framework for social media data[J]. IEEE Trans Knowl Data Eng 26(12):2914–2927
30. WANG C, JIN C (2012) Based on the established vocabulary of Yi automatic segmentation system design and implementation[J]. science technology and. Engineering 10:020
31. Weibo – SINA, in: http://english.sina.com/weibo/
32. Weston J, Ratle F, Mobahi H et al (2012) Deep learning via semi-supervised embedding[J]. Lect Notes Comput Sci 7700:1168–1175
33. Wikipedia, in: http://www.wikipedia.com/
34. Xianghan Z, Nan C, Zheyi C, Chunming R, Guolong C, Wenzhong G (2014) Mobile cloud based framework for remote-resident multimedia discovery and access. J Intern Technol 15(6):1043–1050
35. Xu Z, Lu R, Xiang L et al (2011) Discovering user interest on twitter with a modified author-topic model[C]//web intelligence and intelligent agent technology (WI-IAT), 2011 IEEE/WIC/ACM international conference on. IEEE 1:422–429
36. Yager RR, Reformat MZ (2013) Looking for like-minded individuals in social networks using tagging and E fuzzy sets[J]. IEEE Trans Fuzzy Syst 21(4):672–687
37. Yan Q, Wu L, Zheng L (2013) Social network based microblog user behavior analysis[J]. Phys A: Stat Mech Appl 392(7):1712–1723
38. Yang MS, Lai CY (2012) Lin C Y. a robust EM clustering algorithm for Gaussian mixture models[J]. Pattern Recogn 45(11):3950–3961
39. Yu K, Dang X, Bart H et al (2014) Robust model-based learning via spatial-EM algorithm[J]. Knowl Data Eng IEEE Trans 27(6):1–1

40. Zarrinkalam F, Kahani M, Bagheri E (2018) User interest prediction over future unobserved topics on social networks ☆[J]. Inform Retrie J
41. Zhang Z, Zhou T, Zhang Y (2010) Personalized recommendation via integrated diffusion on user–item–tag tripartite graphs[J]. Phys A: Stat Mech Appl 389:179–186
42. Zheng XH, An DY, Chen X, Guo WZ (2015) Interest Prediction in Social Networks based on Markov Chain Modeling on Clustered Users[J]. Concurr Comput: Pract Exp
43. Zhepeng L, Xiao F, Xue B, Olivia R (2017) S. Utility-based link recommendation for online social networks[J]. Manag Sci 63(6):1938–1952

Xianghan Zheng is an professor in the College of Mathematics and Computer Sciences, Fuzhou University, China, and Fujian Provincial Key Laboratory of finance and technology innovation (Fuzhou University). And he is also an professor in the Key Laboratory of Spatial Data Mining & Information Sharing, Ministry of Education. His current research interests include Big Data, Cloud Computing Services and Applications.

Wenfei Zheng is master student in the College of Mathematics and Computer Sciences, Fuzhou University, China. Her current research interests include Big Data applications, Cloud Computing Services.

Yang Yang is an associate professor in the College of Mathematics and Computer Science, Fuzhou University, China. She received the B.Sc. degree from Xidian University, Xi'an, China, in 2006 and Ph.D. degrees from Xidian University, China, in 2012. Her research interests are in the area of cloud computing security, big data security, and privacy protection.

Wenzhong Guo is an professor in the College of Mathematics and Computer Sciences, Fuzhou University, China, and he is also director of Fujian Provincial Key Laboratory of Networking Computing and Intelligent Information Processing. His research interests include mobile computing, wireless sensor network and evolutionary computation.

Victor Chang is an Associate Professor in Information Management and Information Systems at International Business School Suzhou (IBSS), Xi'an Jiaotong Liverpool University, China. His research interests included Big Data, Cloud Computing, Security and Applications.