Markov Based Social User Interest Prediction

Dongyun An[1,2(✉)] and Xianghan Zheng[1,2]

[1] College of Mathematics and Computer Science,
Fuzhou University, Fuzhou, China
dongyun_an@163.com, xianghan.zheng@fzu.edu.cn
[2] Fujian Key Laboratory of Network Computing and Intelligent Information
Processing, Fuzhou, China

Abstract. In this paper, we propose a new approach to predict users' interest eigenvalues based on multi-Markov chain model, which provides a better personalized service for the users timely. We first collect a dataset from Sina Weibo that includes 4613 users and more than 16 million messages; Then, preprocess data set to obtain users' interest eigenvalues. After that, divide users into several categories and establish multi-Markov chain to predict users' interest eigenvalues. Our experiments show that using multi-Markov model to predict users' interest eigenvalues is feasible and efficient, and could predicting both long-term and short-term user interests based on a suitable selection of the initial state distribution, λ.

Keywords: Social network · Sole–Markov chain · Enhanced–Markov chain · Interest eigen values

1 Introduction

With the development of Internet technology and the emerging forms of media, the Internet entered into a mass of information age [1]. At the same time a new type of information sharing and publishing platform (such as Weibo emergences, so the degree of Internet users' participation and active in China is showing explosive growth. Through this platform, users can express their views by posting some original essays or sharing information [2]. Therefore, effective feature learning and interest prediction [3] is significant not only for users (e.g., looking for users with similar interests [4]), etc.), but also for service providers in a set of application scenarios (e.g., user behavior analysis, personalized recommendation).

Most existing research considers user interest prediction from mainly three aspects: user registration information [5], browsing and posting history [6], and social interaction and relationship [7]. But prediction performance based on these aspects is unsatisfactory. In this paper, we investigate the social network environment and propose a Markov chain model that is feasible and efficient in predicting both long-term and short-term user interests. The contribution of this paper can be concluded that: first, develop specific data crawler to collect dataset from Sina Weibo. After features extraction through a set of operations, each user could be converted and represented as a feature vector; second, obtain user interest eigenvalue sequence by establishing a

A. Bikakis and X. Zheng (Eds.): MIWAI 2015, LNAI 9426, pp. 376–384, 2015.
DOI: 10.1007/978-3-319-26181-2_35

sole–Markov chain model; implement the SOM algorithm to find similarities among users and construct enhanced–Markov chain model that merges users into specific predefined interest categories; finally, conduct experiments to validate the feasibility and efficiency of the proposed solution; validate that the proposed solution can be implemented for both long-term and short-term user interest predictions.

The rest of the paper is organized as follows. Section 2 presents the background of social network and Markov Chain and reviews existing research on user interest prediction. Section 3 introduces dataset preprocess and feature vector extraction. Section 4 describes the construction of sole–Markov chain and enhanced–Markov chain models for user clustering and interest prediction. Experiments and evaluations are conducted in Sect. 5. Finally, conclusions are drawn in Sect. 6.

2 Related Work

In recent years, a lot of research on how to predict user interest has been undertaken. In the view of industry, Twitter and Facebook mention in their reports that they are using AI (deep learning) to understand the significance behind users' posting messages. However, the detailed techniques are proprietary and refuse to public. In academia, related works are: Attenberg et al. [8] propose a user interest prediction mechanism by analyzing the content of messages posted by users or by analyzing their interest eigenvalues. Xu et al. [9] propose a modified author-topic model to discover topics of interest on Twitter by filtering interest-unrelated tweets (noisy posts) from the aggregated user profiles. Yan et al. [10] establish a human dynamic model codriven by interest and social identity and show that user interest in sending posts is positively correlated with the number of comments on their previous posts. In the field of future interest prediction. Nori et al. [11] propose a new graphic representation (Action Graph) for modeling user multinomial with time-evolving actions and predict user interest by computing the similarity between each user and a set of resources. However, this ignores the influence of the user's friends on his or her interests.

In summary, existing research explores user interest from mainly three aspects: user registration information, browsing and posting history, social interaction and relationships. Compared to existing research, our work contains a few distinguished points: (a) investigate and consider the factor of time and influence among friends or similar users, and propose a possible solution that combines Markov model with clustering technology; (b) through the construction of the sole–Markov chain and enhanced–Markov chain models, the proposed solution is capable of providing excellent performance with the true positive rate of clustering reaching 88.5 %.

3 Dataset Collection and Analysis

We develop a specific data crawlers and feature collection mechanisms for dataset collection: Firstly, Hundred normal users (celebrity, company, and government that post, repost, and comment frequently) with 20 interest categories of Weibo messages are manually selected as the data source. Secondly, specific data crawlers are developed

for the ordinary user and for the hot Weibo category. And finally, 4,612 Weibo users, and 20 categories of hot Weibo messages are extracted. After that, for each user, the basic user information (e.g., username) is retrieved by the Weibo API. Through the username, it is possible to retrieve a set of message IDs through which the text messages can be obtained. Finally, more than 16 million messages are crawled. Finally, for each user, a feature vector is constructed according to the crawled user and the message information with the operations in the following section.

As soon as the dataset is crawled, it is preprocessed [12] and each user's messages are converted into a vector which will be used in the establishment of the model.

1. Word Segment. This paper uses the Chinese Institute of Computing segmentation system (ICTCLAS) [13] divided user message content into separated words.
2. Frequency Statistics. The TF-IDF (term frequency–inverse document frequency) [14] algorithm to obtain the keywords appearing in 20 predefined interest categories is used. Finally, the top *50* keywords in each category are extracted. After de-duplication, *579* keywords as user interest eigenvalues are obtained.
3. Feature Vectors Generation. Based on these *579* keywords, it is possible to convert each user's messages into a feature vector.

4 Markov Based User Interest Modeling

Figure 1 illustrates the system framework of proposed interest prediction solution. The concept is: After generation of a series of feature vectors (described in previous section), Markov chain model is implemented to construct the prediction model, and generates a series of user interest categories.

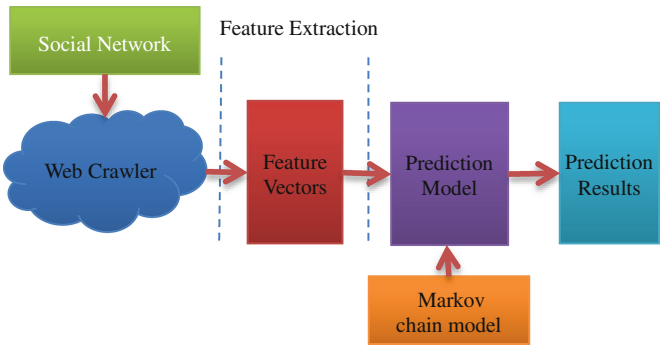

Fig. 1. Overview of interest prediction model

4.1 Sole-Markov Chain Based Interest Prediction (SMC)

According to user interest eigen values, a sole–Markov chain model [15] can be constructed.

A sole–Markov chain model can be represented as a triplet, $MC = <X, A, \lambda>$, where X is a discrete random variable in the range of $\{x_1, x_2, \ldots, x_n\}$, in which each x_i represents user interest eigenvalue. A is the transition rate matrix and λ is the initial state distribution represented as followed:

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

$$\lambda = (p_i) = (p_1, p_2, \ldots, p_n) \tag{2}$$

Where $p_{ij} = P(X_t = x_j \mid X_{t-1} = x_i)$ refers to the transition probability from state x_i to state x_j; $p_i = P(X_{t=0} = x_i)$.

Through the collection of user messages in a certain time period, it is possible to extract a sequence of user interest eigenvalues (variables x). After that, with the maximum likelihood estimation function, it is possible to estimate the value of the parameters in a SMC model, referred to in the following formulas:

$$p_{ij} = \frac{S_{ij}}{\sum_{j=1}^{n} S_{ij}} \qquad p_i = \frac{\sum_{j=1}^{n} S_{ij}}{\sum_{i=1}^{n}\sum_{j=1}^{n} S_{ij}} \tag{3}$$

Where S_{ij} refers to the number of state pairs (x_i, x_j) appearing in users' messages posted.

Let vector $H(t) = [0, 0, \ldots, 1]$ refer to the user interest eigenvalue sequence in time point t and $V(t) = [P(X_t = x_1), P(X_t = x_2), \ldots, P(X_t = x_n)]$ refer to the probability of each eigenvalue. Therefore, it is possible to predict user interest eigenvalue with the formula 4. And the most related user interest eigenvalue is the highest probability value in vector $V(t)$. With the multistage weighted combination model (for considering historical user interest eigenvalues), prediction accuracy can be improved as represented in formulas 5 and 6:

$$V(t) = H(t-1) \times A \tag{4}$$

$$\begin{aligned} V(t) = {} & w_1 H(t-1) \times A^1 + w_2 H(t-2) \times A^2 \\ & + \cdots + w_h H(t-h) \times A^h \end{aligned} \tag{5}$$

$$w_1 + w_2 + \cdots + w_h = 1 \tag{6}$$

Experimental results (Sect. 5.2) show that prediction accuracy increases with the higher value of h, until stabilized eventually.

4.2 Enhanced-Markov Chain Based Interest Prediction (EMC)

Based on the SMC, we further propose an Enhanced-Markov chain based interest prediction. First, two assumptions about user interest eigenvalue sequences are described:

Assume that there are K categories of interests, represented as $C = \{c_1, c_2, \ldots, c_k\}, P(C = c_k)$ refers to the probability of the *i-th* category the user belongs to, then, for each user:

$$\sum_{k=1}^{K} P(C = c_k) = 1 \tag{7}$$

Assume that users in the same interest category have similar behavior features and that the corresponding interest eigenvalues sequences are random process that follow the discrete homogeneous Markov chain.

With above two assumptions, it is possible to construct a user interest prediction classification model containing multiple Markov chains, known as the EMC model.

The EMC interest model is defined as a quaternion: $< X, K, P(C), MC >$, in which X is a discrete random variable in range $\{x_1, x_2, \ldots, x_n\}$, where each x_i represents an interest eigenvalue, $C = \{c_1, c_2, \ldots, c_k\}$ represents a group of user interest categories with the number k, $P(C = c_k)$ refers to the probability of the i-th category the user belongs to, $MC = \{mc_1, mc_1, \ldots, mc_k\}$ expresses a set of Markov chains and each element mc_k is the Markov eigenvalue chain that belongs to a specific category c_k. The transition rate matrix A_k of mc_k and the initial state distribution λ_k could be expressed as:

$$A = (p_{kij}) = \begin{bmatrix} P_{k11} & P_{k12} & \ldots & P_{k1j} & & P_{k1n} \\ P_{k21} & P_{k22} & \ldots & P_{k2j} & & P_{k2n} \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ P_{ki1} & P_{ki2} & \ldots & P_{kij} & \ldots & P_{kin} \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ P_{kn1} & P_{kn2} & \ldots & P_{knj} & \ldots & P_{knn} \end{bmatrix} \tag{8}$$

$$\lambda_k = (p_{ki}) = (p_{k1}, p_{k2}, \ldots, p_{kn}) \tag{9}$$

According to Definition 3, based on user interest eigenvalue sequences, it is possible to construct a set of Markov chains. Formula 10 are expressed for calculating p_{kij}, with p_{ki} belonging to A_k. Where k represents the number of interest categories; S_{kij} represents the number of status pairs (x_i, x_j) appearing in user content; a_{kij} is a super parameter as formula 11 shows:

$$p_{kij} = \frac{S_{kij} + \alpha_{kij}}{\sum_{j=1}^{n} (S_{kij} + \alpha_{kij})} \quad p_{ki} = \frac{\sum_{j=1}^{n} S_{kij} + \alpha_{kij}}{\sum_{i=1}^{n} \sum_{j=1}^{n} (S_{kij} + \alpha_{kij})} \tag{10}$$

$$\alpha_{kij} = \frac{\beta}{n \times n} \tag{11}$$

Where β is the constant value of the problem domain size n.

In cases where the transfer matrixes of two users have a high degree of similarity, δ_{kl}, let the two matrixes merge together, with the calculation formulas:

$$CE(p_{ki}, p_{li}) = \sum_{j=1}^{n} p_{kij} \log \frac{p_{kij}}{p_{lij}} \tag{12}$$

$$Similarity\,(A_k, A_l) = \sum_{i=1}^{n} CE(p_{ki}, p_{li})/n \tag{13}$$

$$\begin{aligned} \delta_{kl} &= Similarity(mc_k, mc_l) \\ &= \frac{2}{Similarity(mc_k, mc_l) + Similarity(mc_l, mc_k)} \end{aligned} \tag{14}$$

Where $CE(p_{ki}, p_{li})$ is the cross entropy of p_{ki} and p_{li}. In case the δ_{kl} value falls between mc_k and mc_l, the Markov chain is large enough or infinite and the corresponding two users are regarded to be in the same interest category, with merging formulas that follow:

$$p_{(k+l)ij} = \frac{S_{kij} + S_{lij} + \alpha_{(k+l)ij}}{\sum_{j=1}^{n} \left(S_{kij} + S_{lij} + \alpha_{(k+l)ij}\right)} \tag{15}$$

$$p_{(k+l)i} = \frac{\sum_{j=1}^{n} \left(S_{kij} + S_{lij} + \alpha_{(k+l)ij}\right)}{\sum_{i=1}^{n} \sum_{j=1}^{n} \left(S_{kij} + S_{lij} + \alpha_{(k+l)ij}\right)} \tag{16}$$

Step by step, user interest prediction with the EMC model can be generated.

5 Experiment Analysis

In this section, an experimental analysis of our proposed solution from four aspects is introduced: a user clustering experiment, prediction comparisons between the SMC and EMC models. Based on the collected dataset containing 4600 users, 3700 users are randomly selected with the Pareto principle as training data and the messages of the remaining 900 users are used as testing data. For simplification, the impact of the events that cause interruption is neglected. The experiment is carried out in MATLAB environment running in two Core i5-3470, 2 * 3.20 GHZ CPU.

5.1 User Clustering

The SOM neural network algorithm [16] is used to cluster users. The SOM algorithm reduces user n-dimensional original transfer matrixes into two-dimensional matrixes and keeps the original topology of the user transfer matrix. From the clustering result shown in Fig. 2, it can be seen that a set of clusters are formed, but these clusters have a lot of noise data. Further investigation reveals that this is caused by a group of spammers who might distribute spam messages in a lot of interest fields. These spam messages greatly reduce prediction accuracy. For noise filtering, the independent component analysis method (provided by Matlab) to remove spammers is imported and denoised user interest clusters are obtained, as shown in Fig. 3.

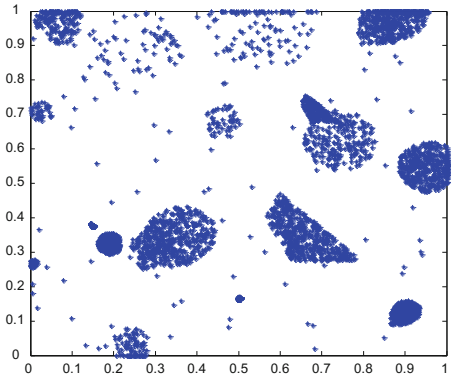

Fig. 2. User classification based on SOM neural network algorithm

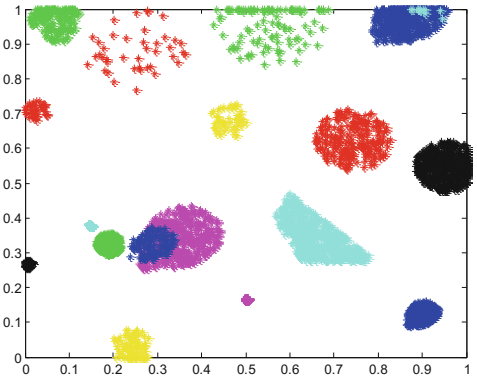

Fig. 3. User classification after denoised

5.2 Prediction Comparisons Between the SMC and EMC Models

In this experiment, the selected user in Sect. 5.1 was reused and the prediction of accuracy between the SMC and EMC models was compared. The results in Fig. 4 show that (1) the prediction is independent of the number of order h (similar to the result of the EMC model described in Sect. 4.2) and (2) the EMC model is capable of separating interest categories with bigger intervals (from 0.03 to 0.35) than the SMC model and, therefore, capable of obtaining the most suitable interest category classification result.

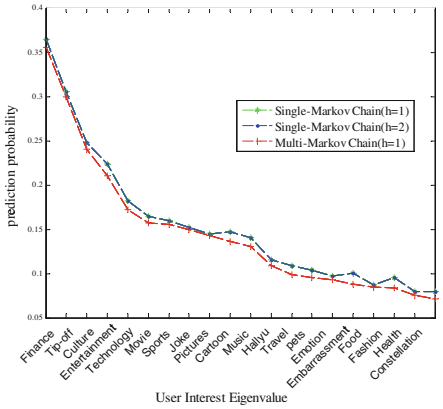

Fig. 4. Prediction accuracy of sole and enhanced-Markov chain model

From the training dataset, 20 users from each category, with 400 users in total, are randomly selected. After that, both SMC and EMC based approaches are implemented, with the average prediction of accuracy results listed in Table 1. The results show that the average value of the SMC model is only 0.5249, with a variance value of 0.0460, while the corresponding values of the EMC model are 0.8699 and 0.0007. This further proves that the EMC based interest prediction model is capable of achieving better accuracy of prediction.

Table 1. Predictive accuracy of sole and enhanced-Markov chain model

Category	SMC	EMC
Average	0.5249	0.8699
Variance	0.0460	0.0007

6 Conclusions

Social user interest prediction has become an important topic in the social network research field. In this paper, interest prediction based on the Markov chain modeling on clustered users is introduced. The solution considers user content feature and obtains user interest eigenvalue sequences to a establish SMC model; implement user clustering algorithms to construct a EMC model that classifies different users into specific predefined interest categories. Through a multitude of analyses, experiments and evaluations, it can be concluded that the proposed solution is feasible, efficient, and

capable of achieving a much higher accuracy of prediction than any of the other existing approaches.

Acknowledgements. The authors would like to thank the support of the Technology Innovation Platform Project of Fujian Province under Grant No. 2009J1007, the Program of Fujian Key Project under Grant No. 2013H6011, the Natural Science Foundation of Fujian Province under Grant No. 2013J01228.

References

1. Wasserman, S.: Social network analysis: Methods and applications. Cambridge University Press, Cambridge (1994)
2. Statista, in: http://www.statista.com/
3. Bao, H., Li, Q., Liao, S.S., et al.: A new temporal and social PMF-based method to predict users' interests in micro-blogging. Decis. Support Syst. **55**(3), 698–709 (2013)
4. Chen, K.H., Han, P.P., Wu, J.: User clustering based social network recommendation. Jisuanji Xuebao(Chinese Journal of Computers) **36**(2), 349–359 (2013)
5. Yang, S.H., Long, B., Smola, A., et al.: Like like alike: joint friendship and interest propagation in social networks. In: Proceedings of the 20th International Conference on World Wide Web, pp. 537–546. ACM (2011)
6. Van Iddekinge, C.H., Putka, D.J., Campbell, J.P.: Reconsidering vocational interests for personnel selection: The validity of an interest-based selection test in relation to job knowledge, job performance, and continuance intentions. J. Appl. Psychol. **96**(1), 13 (2011)
7. La Greca, A.M., Harrison, H.M.: Adolescent peer relations, friendships, and romantic relationships: Do they predict social anxiety and depression? J. Clin. Child Adolesc. Psychol. **34**(1), 49–61 (2005)
8. Attenberg, J., Pandey, S., Suel, T.: Modeling and predicting user behavior in sponsored search. In: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1067–1076. ACM (2009)
9. Xu, Z., Lu, R., Xiang, L., et al.: Discovering user interest on twitter with a modified author-topic model In: 2011 IEEE/WIC/ACM International Conference on IEEE Web Intelligence and Intelligent Agent Technology (WI-IAT), vol. 1, pp. 422–429 (2011)
10. Yan, Q., Yi, L., Wu, L.: Human dynamic model co-driven by interest and social identity in the Microblog community. Physica A **391**(4), 1540–1545 (2012)
11. Nori, N., Bollegala, D., Ishizuka, M.: Interest prediction on multinomial, time-evolving social graph. In: IJCAI, vol. 11, pp. 2507–2512 (2011)
12. Phan, X.H., Nguyen, C.T., Le, D.T., et al.: A hidden topic-based framework toward building applications with short Web documents. IEEE Trans. Knowl. Data Eng. **23**(7), 961–976 (2011)
13. Wang, C., Jin, C.: Based on the established vocabulary of yi automatic segmentation system design and implementation. Sci. Technol. Eng. **10**, 020 (2012)
14. Teevan, J., Ramage, D., Morris, M.R.: # TwitterSearch: a comparison of microblog search and web search. In: Proceedings of the Fourth ACM International Conference on Web Search and Data Mining, pp. 35–44. ACM (2011)
15. Liu, C.: Stochastic Process (fourth edition). Huazhong University of Science and Technology Press, Wuchang Yu Jiashan, vol. 8, pp. 1–113 (2008)
16. Ghaseminezhad, M.H., Karami, A.: A novel self-organizing map (SOM) neural network for discrete groups of data clustering. Appl. Soft Comput. **11**(4), 3771–3778 (2011)