Link Prediction Model for Anchor Chain Connection Method

Shou-meng Huang[1, 2]

[1]College of information and intelligence engineering, Sanya University, Sanya, China
[2]Academician Guoliang Chen Team Innovation Center, University of Sanya, Sanya, China
Huang123888@126.com

Abstract:
In order to improve the accuracy of the location-based social network (LBSN) multi-source heterogeneous data link prediction, a model named anchor link-predict (AL-P), based on the anchor link method is proposed. Firstly, the network representation learning method is used to obtain the user relationship topology in LBSN. The matrix decomposition method is used to obtain the user sign-in record representation space in LBSN. Then, the anchor link method is joined the user relationship topology and user check-in record, and the potential relationship between them is mined. Finally, the experimental analysis shows that the Al-P model can significantly improve the link prediction effect under different evaluation indexes compared with the existing predicting models of same type.

Keywords: *anchor link; link prediction; LBSN; prediction algorithm*

1 INTRODUCTION

In the context of big data, social networks have multidimensional and relevant data. It is worth thinking how to make full use of these features. Therefore, it is necessary to build a mathematical model to mine the potential relationship information between these data, such as location-based Social Network (LBSN), which generally contains two kinds of data [12]: One is the data of users themselves in the network, and the other is the relational data existing between users. The temporal and spatial information between users can be mined from LBSN for various application activities [6], such as friend recommendation [8], interest recommendation [2], trajectory recovery [14], behavior prediction [15] and other application scenarios. In fact, in the analysis of social networks, link prediction has always been a research focus of information recommendation system, which is to find out the information that may exist node link from the known node information in the network. As far as social networks are concerned, in the field of link prediction, information recommendation can be realized based on user check-in records and social network data. Li Zhepeng et al. [12] proposed the prediction method of the fusion of these two heterogeneous data in LBSN, which utilized network representation learning and matrix decomposition of data node information, and completed the prediction task through Mosaic and fusion scheme. Hu Wei et al. [6] constructed a link model by integrating network users' time behavior with social theory using probabilistic methods. Li Jichao et al. [9] used the association information between network structure and node formation time to construct a prediction method, while Liang Wenxin et al. [10] quantified the association between user node social graph attributes, user profile information and mobile features to model and predict friend relationships. The joint prediction model is constructed by combining user topic features and node topological structure. Ling Jiang et al. [7] studied the network crossover view and completed the link prediction task by learning link representation and user attribute representation. Complex network link prediction tasks are also completed from different perspectives by heterogeneous data sources for mixed prediction [18], based on cluster analysis [16], from the perspective of informatics analysis [17], based on local similarity of nodes [1], based on matching degree of resource transmission [11], based on mutual information [13]. In order to further improve the link performance and forecast accuracy, this study proposed an anchor link scheme to replace the data in the literature [12] fusion splicing solutions, deeper mining LBSN link information of each node and its associated data is two different anchor link node node for the public, the mapping relationship they performed by neural network algorithm.

B. Fox et al. (Eds.): IC-ICAIE 2022, AHCS 9, pp. 657-663, 2023.
https://doi.org/10.2991/978-94-6463-040-4_100

2 PREVIOUS WORK

2.1 *Network represents learning*

There are many methods to extract information of network nodes, among which the most popular is network representation learning method [3], which uses deep learning technology to quantify the dimensionality reduction process of network space, so as to extract feature vectors of each node in the network. As shown in Figure 1, Deep Walk [9] implements the general process of network representation learning to obtain user node vectorization.

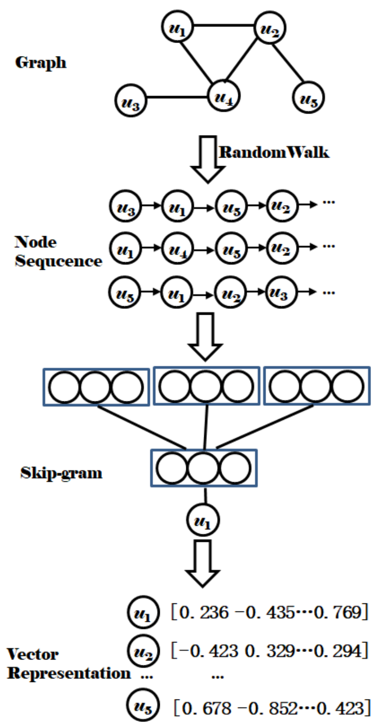

Figure 1 Network represents the learning process

At present, in the field of friend recommendation of LBSN, network representation learning can be used to complete tasks such as vector representation of low-dimensional space, node classification and clustering, etc. By mining potential relationship rules and features in the network topology of social users, LBSN network relations can be modeled. The user association information of each user node and each node in the LBSN social relationship topology is obtained by vector representation.

2.2 *Matrix decomposition*

Matrix decomposition is actually a matrix transformation method. In order to accomplish the dimensionality reduction task of the matrix, the original matrix is split into multiple associated matrices [4]. In recommendation systems, matrix decomposition is often used to filter certain data to complete recommendation tasks [5]. For example, if the user scores the movie, assume that all the data are obtained from the movie ticket purchasing system and the user rating table is constructed, as shown in Figure 2. The row value is the user ID, the column value is the movie ID, and the symbol is "?", which represents the missing scoring value, that is, the phenomenon that a user does not score a movie without watching it. The next step is to predict the missing score values (assuming a score of 1 ~ 5), the scale of the matrix decomposition of matrix and the product of the relationship between film matrix for the user. And it is hoped to multiply the results with the original score as close as possible of the matrix loss function (conditions), through constant iterative convergence model which can be acquired after several rounds of training optimization, that is, the loss function. Thus, the missing value in the original matrix is replaced by the new matrix obtained by multiplication, that is, the value of the user's interest in unwatched movies is predicted.

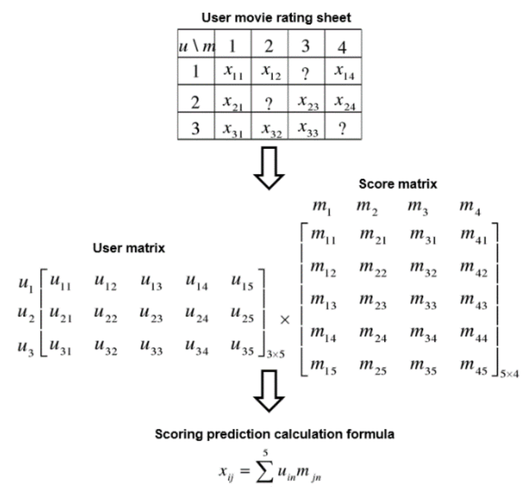

Figure 2 Film review chart of matrix decomposition

2.3 *Relevant prediction models*

Walk2friends [13] model only extracts mobile features of users; The DeepWalk [9] model extracts the user node sequence by walking; LINE [7] model adopts node co-occurrence and conditional probability modeling. GraRep [6] model constructs transfer matrix to extract node similarity; Node2vec [16] Used local and global network attributes to extract node domain; Struc2vec [10] model captures the vectorization modeling of node structure. In this study, node information (user check-in record) and node structure relationship (matrix representation) are used to model the potential relationship, and anchor chain algorithm is used to fuse the two data to vectorize the point-to-point relationship, and the prediction performance is optimized through training set and test set.

3 PREDICTION METHODS

3.1 *Basic definitions*

As for any two social networks $G^v = (U^v, E^v)$ and $G^n = (U^n, E^n)$, the traditional anchoring link method directly predicts whether two users $u_i \cup U^v$ and $u_j \cup U^n$ the same person (that is, $u_i = u_j$), the al-P model constructed in this paper adjusts the vector space of the two users in two LBSN, namely, the anchoring link satisfies the binary discriminating function ϕ :

$$U^v \times U^n \to \{0,1\} \tag{1}$$

whose corresponding formula is

$$\left[(u_i)_{ENC}, (u_j)_{ENC}\right]_{DEC} = \left(u_i \in U^v, u_j \in U^n{}_j\right)_{DEC} \approx P(u_i, u_j), \tag{2}$$

Where: $(\ )_{DEC}$ represents the vector coding of user node, namely the quantized value of node similarity; $(\ )_{ENC}$ Represents the maximum dimension preserving network attribute, that is, node vectorization dimension value; $P(\,)$ is a user-defined similarity measure between nodes.

3.2 *Prediction model*

The anchor link prediction model AL-P mainly has three parts: user access preference learning, update access preference vector and social relationship vectionization. The detailed design architecture is shown in Figure 3. The input values for the user to sign in record and network structure relations, the output for the updated good user access preference vectors, anchor link algorithm is based on three layers neural network model, it receives the user access preference vectors u_i^v and vector u_i^n social network users, will serve u_i^v as a feature data, u_i^n as the corresponding labels make a training on (features, tag), namely (u_i^v, u_i^n) generated and updated user access preference vector u_i^v after completing the training convergence of the anchor chain model. The integrated mapping between layers has the mapping function ϕ and the learning cosine function F.

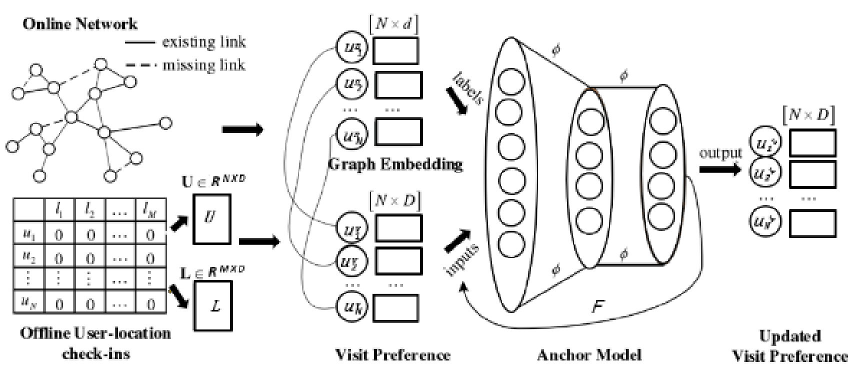

Figure 3 Al-P prediction model

3.3 *Anchor chain connection algorithm*

The core work of the AL-P cabling model is the cabling algorithm, which has two main tasks: calculating the mapping function ϕ and learning the cosine function F. The mapping function ϕ is defined :

$$G^v \to G^n \tag{3}$$

the mapping between the user check-in record and the user-relational topology, with a pair of anchor nodes (u_i, u_j) satisfying $\phi(u_i^v) = u_j^n$. In order to optimize the vector "offset" generated by learning cosine function F to obtain approximate mapping function after continuous training and learning, if three users u_i, u_p, $u_q \in U^v$, u_i and u_p are friends, but u_i and u_q are not friends, the cosine value of vector is satisfied

$$\cos(u_i^v, u_p^v) > \cos(u_i^v, u_q^v) \tag{4}$$

which is the potential association of user vector alignment method to capture nodes, the specific algorithm pseudo-code is as follows.

Input: two heterogeneous networks G^v and G^n; pre-training functions ϕ and F; Parameters w and b. tagged anchor chain assembly E^v.

Output: updated user access preference vector $u_i'^v$.

1: repeat

2: for each epoch do

3: for i = 1 to N do

4: random sampling of a certain number of users $u^{vi} \in U^{v}$ to build a mini-batch

5: calculate $u^{'vi}$ according to ϕ (when i= 1, $\phi' = \phi$)

6: update parameters w and b

7: end for

8: calculate $y = F_{\text{true}}$ and $a = \sigma(F_{predicted})$

9: update parameters w and b again

10: end for

11: until convergence

12: return $\{u^{'vi}\}$

4 EXPERIMENTAL RESULTS AND ANALYSIS

4.1 *Preparation for the experiment*

This experiment uses two different open source datassets Gowalla and Foursquare [7] (among @NY represents New York, @TY represents Tokyo, @WHG represents Washington, and @CCG represents Chicago) based on LBSN. As shown in Table 1, user relationship topology and user check-in records are extracted from them. As well as preprocessing related data, the results are shown in Table 2 below.

Table 1 Experimental data set

Data set	Social network users	Social network link number	Sign in record number	Sign in to users	POI number	Check-in time range
Foursquare@NY	41 408	131 467	227 428	1 085	38 338	12:00-15:00(weekday) 19:00-24:00(weekend) 04/2021-09/2021
Foursquare@TY	80 774	263 684	573 705	2 295	61 858	
Gowalla@WHG	4 456	24 544	158 616	6 316	6 316	12:00-15:00(weekday) 19:00-24:00(weekend) 05/2020-10/2020
Gowalla@CCG	4 219	17 842	162 166	6 249	6 249	

Table 2 Data pretreatment results

Data set	Common number of users	Number of check-in records	Number of samples in test set	Number of training set samples	POI number
Foursquare@NY	588	22 565	1 164	4 648	1 993
Foursquare@TY	1 057	38 744	1 926	7 701	2 214
Gowalla@WHG	881	13 596	1 166	4 663	4 797
Gowalla@CCG	629	10 316	508	2 035	3 269

The reference models of this experiment include Walk2Friends [14] and DeepWalk [15], where in Walk2Friends extracts user movement features, while DeepWalk extracts user node sequences. The operation methods of the two models are unified by random selection of Average. The evaluation indexes are AUC, accuracy, recall and F1 value (harmonic average of accuracy and recall).

4.2 *Experimental results*

The evaluation index AUC can detect the performance of the link prediction model, and it represents the probability value of positive/negative samples. The experimental results are shown in Table 3, indicating that in different data sets, the AUC value of AL-P model is superior to that of existing models walk2Friends, DeepWalk, LINE, GraRep, Node2vec and struc2vec, where data fusion operation uniformly adopts Average operation.

Table 3 AUC prediction results of AL-P

Methods	@NY	@TY	@WHG	@CCG
walk2friends	0.542 1	0.539 8	0.460 2	0.433 7
DeepWalk	0.653 3	0.625 1	0.711 3	0.609 2
LINE	0.640 9	0.594 1	0.647 2	0.631 9
GraRep	0.646 8	0.645 3	0.662 1	0.578 8
node2vec	0.638 7	0.651 2	0.693 7	0.586 6
struc2vec	0.754 4	0.752 5	0.796 6	0.689 8
AL-P	0.932 7	0.922 3	0.857 2	0.822 1

For Foursquare data sets (@NY and @TY) and Gowalla data sets (@WHG and @CCG), the measurement experiment of evaluation index accuracy, recall rate and F1 value is completed. Table 4 is the prediction result of Foursquare data set, And Table 5 is the prediction result of Gowalla data set. It can be seen from Table 4 and Table 5 that in the link prediction task, the AL-P prediction effect is the best in the two data sets. This is because the AL-P model includes user check-in preference information, which makes social network information more indicative than user movement information and effectively improves the accuracy of link prediction. However, the poor result of Walk2Friends is due to the lack of user network structure information during the prediction.

Table 4 The predicted results of the Foursquare dataset

Prediction method	@NY Data set			@TY Data set		
	Precision	Recall	*F1 value*	Precision	Recall	*F1 value*
walk2friends	0.535 8	0.607 7	0.569 9	0.524 1	0.668 5	0.587 4
DeepWalk	0.611 4	0.533 6	0.563 6	0.579 3	0.672 6	0.622 5
LINE	0.620 6	0.595 6	0.607 7	0.569 5	0.554 2	0.561 8
GraRep	0.601 6	0.600 3	0.590 2	0.394 2	0.416 1	0.405 1
node2vec	0.612 6	0.501 1	0.546 2	0.586 1	0.667 1	0.623 4
struc2vec	0.738 1	0.591 2	0.657 4	0.732 5	0.581 1	0.647 7
AL-P	0.868 9	0.911 2	0.890 3	0.820 8	0.875 4	0.847 2

Table 5 Prediction results of Gowalla dataset

Prediction method	@WHG Data set			@CCG Data set		
	Precision	Recall	*F1 value*	Precision	Recall	*F1 value*
walk2friends	0.542 6	0.515 6	0.528 7	0.591 1	0.460 7	0.517 8
DeepWalk	0.811 7	0.458 9	0.586 3	0.826 6	0.606 2	0.465 8
LINE	0.719 4	0.418 8	0.522 4	0.450 2	0.726 2	0.552 2
GraRep	0.813 2	0.440 5	0.571 8	0.800 2	0.407 4	0.443 6
node2vec	0.786 9	0.443 6	0.567 4	0.781 4	0.323 2	0.457 1
struc2vec	0.772 3	0.558 2	0.667 9	0.748 6	0.435 2	0.548 7
AL-P	0.826 2	0.901 4	0.864 9	0.850 3	0.767 7	0.784 7

5 CONCLUSIONS

In order to improve the effect of link prediction, this study proposes an anchor chain method, which completes the mapping of two heterogeneous Spaces through user vector alignment, obtains new user access preference vector, and then updates the social relationship vector after training fusion, so as to further improve the comprehensive performance of link prediction. It still needs to be further studied on how to learn better user representation from mobile data and social data, and mine the potential value of relevant user data. If we model the spatio-temporal behavior, we can better complete the prediction task.

ACKNOWLEDGEMENTS

It is supported by Hainan Provincial Natural Science Foundation of China (622RC735);

It is supported by Hainan Provincial Philosophy and Social Sciences Foundation of China (HNSK(ZC)21-109);

Project supported by the Education Department of Hainan Province, project number： Hnky2021-51

REFERENCES

[1] Bai H, Ma Y.L, BI Y, et al.(2020). A Complicated Network Link Prediction algorithm based on Local Similarity of Nodes.J. Computer Applications and Software. 37(05),298-301.

[2] Daud A, Ahmad M, Malik M.S.I, et al. (2015).Using Machine Learning Techniques for Rising Star Prediction in Co-author Network.J. Scientometrics.2,1570-1583

[3] Huang S.M, Xia W.X(2021). Link Prediction Model Based on Anchor Link Method in LBSN .J. Journal of Hainan Tropical Ocean University. 28(05), 72-77.

[4] Huang S.M, Xia W.X.(2021). A link prediction method based on LSH technology .J. Information Recording Materials.22(07),139-142.

[5] Huang S.M, (2021). Heterogeneous Network Link Prediction Model Based on Supervised Learning.J. Computer Science.48(S2),111-116.

[6] Hu W, Li J, Cheng J, et al. (2020). Security Monitoring of Heterogeneous Networks for Big Data Based on Distributed Association Algorithm.J. Computer Communications. 5,206-217

[7] Jiang L, Yang C.C. (2017). User Recommendation in Healthcare Social Media by Assessing User Similar-ity in Heterogeneous Network.J. Artificial Intelligence in Medicine. 9,211-223

[8] Kovacs I.A, Luck K, Spirohn K, et al.(2019). Network-based Prediction of Protein Interactions.J.Nat-ure Communications.1,107-119

[9] Li J, Zhao D, Ge B.F, et al. (2018). A Link Prediction Method for Heterogeneous Networks Based on BP Neural Network.J. Physica A-Statistical Mechanics and Its Applications. 495(1),10-18

[10] Liang W.X, LI X, He X.S,et al. (2018).Supervised Ranking Framework for Relationship Prediction in He-terogeneous Information Networks.J. Applied Intelligence. 5,1111-1123

[11] Liu S.X, LI X, Chen H.C., et al.(2020). Link prediction method based on matching degree of resource transmission for complex network.J. Journal of Communications. 41(06),70-79.

[12] Li Z.P, Fang X, Bai X, et al. (2016). Utility-based link recommendation for online social networks.J. Management Science.6,1938-1947

[13] Qi F.P, Wang T, FU Z.Q.(2020). Link prediction in complex networks based on mutual information .J. Journal of University of Science and Technology of China. 50(01),57-63.

[14] Shi C, Li Y.T, Zhang J.W, et al.(2016). A Survey of Heterogeneous Information Network Analysis.J. IEEE Transactions on Knowledge and Data Engineering.1,108-119

[15] Sun Y.Z, Han J.W, Yan X.F, et al(2021).. Pathsim: Meta path-based Top-k Similarity Search in Hetero-geneous Information Networks.J. Proceedings of the VLDB Endowment.11,120-129

[16] To J, Sheng H.L.(2019). Research on Link Prediction Performance of Complex Networks Based on Clustering Analysis. J. Computing Technology and Automation, 4,144-150.

[17] Wang H, Le Z.C, Gong X, et al. (2020). Link Prediction of Complex Networks is Analyzed from the Perspective of Informatics. J. Journal of Chinese Computer Systems. 41(02),316-326.

[18] Wu Bangying. (2019). The Research on Link Prediction Problem Based On Unifying Multisource And Heterogenous Data, Chengdu: University of Electronic Technology of Chin