Dynamic Community Detection Algorithm Based On Hidden Markov Model

Zhe Dong

NO.393, Qinyuan street, Jiyuan, Henan, China

zdasika@163.com

Keywords: Dynamic Social Network; Hidden Markov Model; Community Structure

Abstract. The HMM_DC algorithm is proposed based on the Hidden Markov Model to detect the community in dynamic social network. The algorithm transforms the community detection problem to get the optimal status chain in Hidden Markov Model with considering the history information and characteristics in dynamic social network. The algorithm uses the observed chain and status chain to represent the community structure and node information and can identify the community structure without extra information. The experiment results show that HMM_DC algorithm is available and performs effectively and accurately in identifying the community structure in the dynamic social network and the value of Q and NMI can raise 28% and 20% at least.

Introduction

More and more new applications play important roles on social network with the development of Internet. SNA(Social Network Analysis) has been the main method to study the structure and characteristic of the network system. With the research of many complex network, community are widespread in social network[1,2]. The nodes have common properties and contact frequently in the community and contact fewer between communities.Detecting and studying the community is important to have a good knowledge of the inner structure and analyze the unknown information of social networks. More and more algorithms are proposed to detect the community structure in social network, such as LPA[3] and modularity based algorithm[4,5]. It's worth noting that most of the researches are based on the static social network. However, researchers have found that most of the social networks are dynamic and the individual in the network increase or decrease with the time lapse(Fig.1). Therefore, it is very important to study the community to understand the structure and function in the dynamic network.

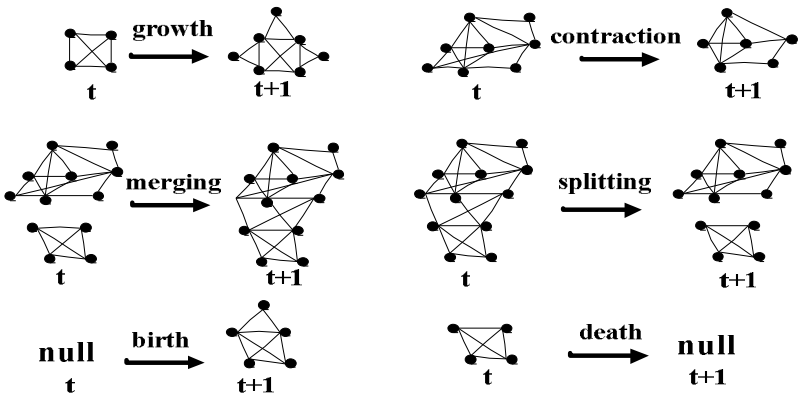

Fig. 1 Community in the dynamic social networks

There are many achievements on detecting community structures in dynamic social networks. Berger-wolf[7] proposed a widespread framework which considered that the dynamic social network could be divided into pieces of static networks of different snapshot. A Social Cost model was proposed to analyze the community in dynamic networks which only consider the individuals in the network. Sun et al[9] gave a new algorithm GraphScope based on information theory, which can detect the community and variety nodes without entering any parameters in order to decrease the

complexity.[10-14] proposed the other remarkable algorithms to detect the community in social network. However, communities in real networks change over time. In contrast to the previous work on tracking independent communities, the evolution of structure within social network should be considered. The structure in current time can be related with the previous structure of the social networks.

Considering the problems in the previous dynamic algorithms, the HMM_DC algorithm is proposed based on the Hidden Markov Model to detect the community in dynamic social network. The algorithm transforms the community detection problem to get the optimal status chain in Hidden Markov Model with considering the history information and characteristics in dynamic social network. The algorithm uses the observed chain and status chain to represent the community structure and node information and can identify the community structure without extra information. The result of the experiments shows that HMM_DC algorithm is available and performs effectively and accurately in identifying the community structure in the dynamic social network and the value of Q and NMI can raise 28% and 20% at least.

Modeling

HMM (Hidden Markov Model, Fig. 2) is a doubly stochastic process and a probabilistic model to describe statistical properties of random processes. It consists of two parts: Markov chains and stochastic process. Markov chain used to be described by the state of the shift and described by the transition probability. General stochastic process used to be described by the relationship between the state and observe sequences and described by the observe probability.

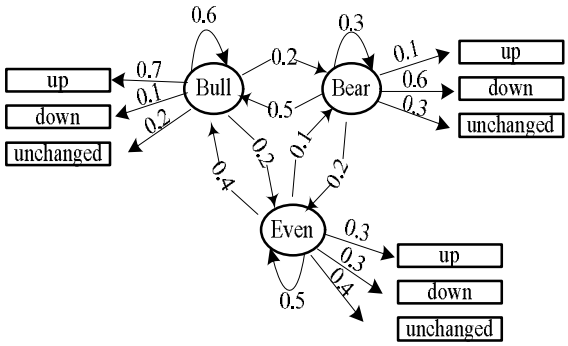

Fig. 2 HMM model[15]

The dynamic social networks are divided into multiple snapshots by time granularity and can be represented by multiple static networks.

We assume that the adjacent networks are time-related and can be described with the first-order Markov Chain. The community structure of the social networks cannot be observed directly, therefore we treated the community in social network as status chain with considering the HMM. The degree distribution of each nodes can be described by observed chain. The algorithm transforms the community detection problem to get the optimal status chain in Hidden Markov Model with considering the history information and characteristics in dynamic social network when compared with the current dynamic community structure detection algorithms. Therefore, HMM_DC algorithm was proposed to detect the community structure in social network with considering the characteristics of HMM and dynamic social networks.

In order to get the optimal status sequence of HMM, we use the Viterbi algorithm which is a dynamic programming algorithm. Viterbi algorithm can find the hidden status of the observed sequence $S_1, S_2, S_3, \mathbf{L}$ and the maximum probability of the hidden status sequence with giving a observed sequence $O_1, O_2, O_3, \mathbf{L}$.

Description of HMM_DC algorithm

Social networks can be described by Graph and expressed by $G = (V, E)$.

V : Collection of nodes in social network
|V|: The total number of nodes in social network
E : Collection of edges in social network
|E| : The total number of edges in social network

Definition 1 (Graph Sequence): Given a graph sequence GS_i which describe the relationship among nodes in social network at one time. So, the dynamic social network G can be represented by collections of graph sequences. $G=(GS_1,GS_2,GS_3,\mathbf{L},GS_t,\mathbf{L},GS_T)$, $1\le t\le T$, $\bigcup_{1\le i\le T} GS_i = G$.

Definition 2 (Status Sequence): The status sequence C in HMM can be described by the community in snapshot in the dynamic social network. Such as $C=(C_1,C_2,C_3,...,C_i,...,C_{Nc})$, N_C is the number of status and q_t is the status in snapshot t.

Definition 3 (Observed Sequence): The observed sequence O in HMM can be described by the nodes' degree in dynamic social network. $O=(d_1,d_2,d_3,\mathbf{L},d_i,\mathbf{L},d_M)$, M is the number of the different observed value, d_t is the degree of the node.

Basically, the HMM can be described by $l=(p,A,B)$ which $p=\{p_i\}$ is the initial probability distribution, $p_i=P(q_1=C_i),1\le i\le Nc$, $\sum_{1\le i\le N} p_i=1$. $A=\{a_{ij}\}$ is the status transition matrix in which the probability of status C_i in t moment and status C_j in $t+1$ moment $B=\{b_j(k)\}$ is the observed probability matrix of status j in which

$$a_{ij}=P\{q_{t+1}=C_j \mid q_t=C_i\},1\le i,j\le Nc \tag{1}$$

$$b_j(k)=P(O_t=d_k \mid q_t=C_j),1\le j\le Nc,1\le k\le M \tag{2}$$

The *Core Nodes* and *status transition probability* are given in order to get A and B, $Sim(x,y)$ is the similarity of vector $x=(x_1,x_2,x_3,\mathbf{L},x_n), y=(y_1,y_2,y_3,\mathbf{L},y_n)$.

$$Sim(x,y)=\sum_{i=1}^{n} x_i\cdot y_i \Bigg/ \left(\sum_{i=1}^{n} x_i^2\cdot\sum_{i=1}^{n} y_i^2\right)^{1/2} \tag{3}$$

Definition 4 (Core Nodes): $NCore_{C_i}$ is the collection of core nodes in C_i, then $Degree(x_i)$ is the degree of node x_i, and $C=(C_1,C_2,C_3,\mathbf{L},C_i,\mathbf{L},C_{Nc})$, $\forall C_i,1\le i\le N_C$.

$$NCore_{C_i}=\arg\max_{x_i\in C_i}(Degree(x_i)) \tag{4}$$

Definition 5 (Status Transition Probability): p_trans_state is the probability of the status of node i transfer to the status C_i.

$$p\_trans\_state=\frac{Sim(i,NCore_{C_i})}{\sum_{k\in(1,Nc)} Sim(i,NCore_{C_k})},i\in V \tag{5}$$

We give the definition of the affiliation function $Com(i,C)$ [7] which represents the degree of the node i in adjacent matrix R belong to the group C. r_{ij} is the member of matrix R and $SumR=\sum_{i,j} r_{ij}$, $r_{i*}=\sum_k r_{ik}$, $r_{*j}=\sum_k r_{kj}$.

$$Com(i,C)=\frac{1}{SumR}\left(r_{ii}+\sum_{j\in C}(r_{ij}+r_{ji})-\frac{r_{i*}r_{*i}}{SumR}\right) \tag{6}$$

Definition 6 (Observed Probability): The observed probability in matrix B is

$$p\_observe=\frac{Com(i,C_k)}{\sum_{j\in(1,Nc)} Com(i,C_j)},i\in V \tag{7}$$

The basic idea of the HMM_DC algorithm is that we assume the current structure is under the influence of previous networks. Then, the algorithm transforms the community detection problem to get the optimal status chain in Hidden Markov Model with considering the history information and characteristics in dynamic social network. The algorithm uses the observed chain and status chain to represent the community structure and node information. Fig. 3 is the framework of HMM_DC algorithm.

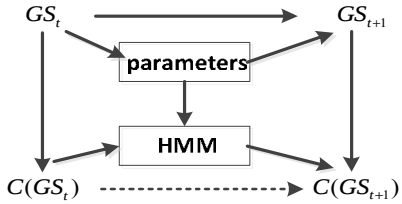

Fig. 3 the framework of HMM_DC algorithm

In this section, we describe the detailed steps of HMM_DC algorithm in Table 1.

In Table 1 we can get that n is the number of nodes and m is the number of edges and N_c is the number of initialization community. In table 1, the complexity is $O(n)$ from line 2 to line 3, and the complexity of line 5 to 7 is $O(n)$ for the algorithm needs to meet every node in network. The complexity of the algorithm is $O(n)+O(n^2 \cdot N_c)+O(n^2 \log n)$ to sum up in conclusion. The complexity of the algorithm is $O(n^2 \cdot N_c)$ in which the number of community is similarity with the number of nodes in uniformity networks.

Table 1 HMM_DC algorithm

Algorithm. HMM_DC.
Input : GS_t 、 GS_{t+1} , Initial community C_ini
Output: the optimal community C_{t+1} in t+1
1. Initialization：len=length(GS_t) , NBC=zeros(len, N_c), $\Pi = p\_initial$ = Rand(1, N_c);
2. for $1 \le i \le len$
3. *NBC(i, C(i))* = 1;
4. end
5. for $1 \le j \le N_c$
6. Find *Ncore* of every community
7. end
8. for $1 \le i1 \le len$
9. for $1 \le j1 \le N_c$
10. Get the similarity between node $i1$ and *Ncore Sim*($i1$, *NCore*);
11. end
12. Get p_trans_state and $\boldsymbol{A}$;
13. for $1 \le k1 \le N_c$
14. Get the affiliation of node $i1$ and C_{k1}
15. end
16. Get $p\_observe$ and $\boldsymbol{B}$;

17. VITERBI(i, π, p_trans_state, $p\_observe$);
18. if P = max(VITERBI) (P is the biggest probability)
19. Put node i in community C that can get the maximum value of P
20. end
21. end
22. Get the optimal C

Experiments

In this section, we compare the HMM_DC algorithm with CDBIA[16]、QCA[11]、MIEN[17]. CDBIA is an increment algorithm. QCA is a self-adaptive algorithm without any parameters. MIEN uses the idea of compress and uncompressing to describe the information in network. The simulation environment is 1G, 3.2GHz, P4 processor and Matlab R2008b integrated environment.

In order to verify the effectiveness of the algorithm, we use the VAST and Enron and Facebook social network datasets to simulate. The VAST dataset is an open competition dataset that come from IEEE VAST 2008 and contains a 400-member group of calling data in ten days. ENRON dataset consists of approximately 1.5 million email communications sent or received by employees in Enron, Inc.. Facebook social network [18] dataset is a collection of friends relationship in Facebook.

In order to evaluate the effectiveness of the algorithm, we use the NMI function and modularity Q function. NMI is used to describe the similarity between detective community and the real community. Then

$$NMI(A,B)=\frac{-2\sum_{i=1}^{c_A}\sum_{j=1}^{c_B}N_{ij}\log\left(\frac{N_{ij}N}{N_{i\bullet}N_{\bullet j}}\right)}{\sum_{i=1}^{c_A}N_{i\bullet}\log\left(\frac{N_{i\bullet}}{N}\right)+\sum_{j=1}^{c_B}N_{\bullet j}\log\left(\frac{N_{\bullet j}}{N}\right)} \tag{9}$$

A and B are the partitioning results in network, N is a confusion matrix and N_{ij} is the number of nodes exist in community i and j simultaneously, $N_{i\bullet}$ is the sum of row i in matrix $\boldsymbol{N}$, $N_{\bullet j}$ is the sum of column j in matrix $\boldsymbol{N}$, C_A and C_B are the number of real community.

Modularity Q function is an important measurement that proposed by Newman and Girvan.

$$Q=\frac{1}{2m}\sum_{uw}\left[A_{uw}-\frac{k_u k_w}{2m}\right]d(c_u,c_w) \tag{10}$$

m is the number of edges in network, k_u and k_w are the degree of node u and w, A_{uw} is the element in adjacent matrix $\boldsymbol{A}$.

VAST dataset: We divide the network by days that we can get ten Graph Sequences.

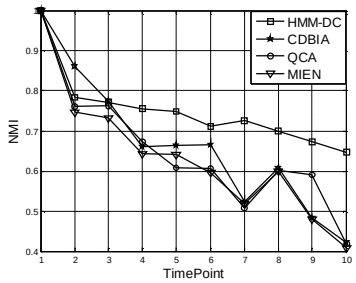

Fig.4 NMI

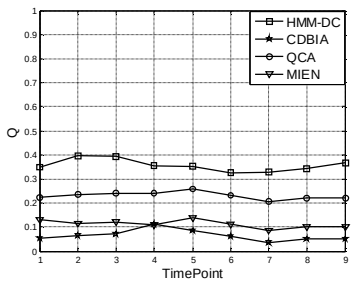

Fig.5 Modularity Q

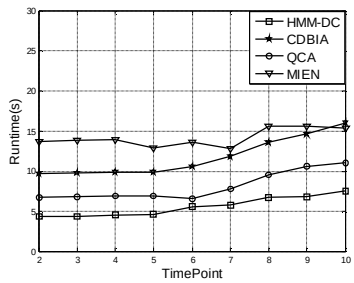

Fig.6 Runtime

Typical community dynamic examples discovered in VAST dataset are shown in Fig. 4–6. In Fig.4, the performance improved 40%, although the NMI decreased 0.08 in the second timepoint. In Fig.5-6

HMM_DC algorithm has an apparently advantage over the other algorithms in modularity and is more efficient than the others algorithms.

ENRON dataset: We divide the network by month and typical community dynamic examples discovered in ENRON dataset are shown in Fig. 7-9.

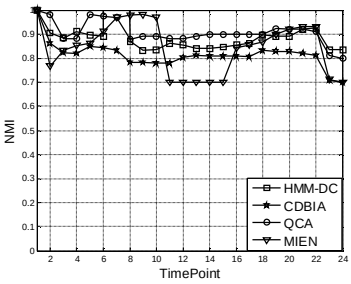

Fig.7 NMI

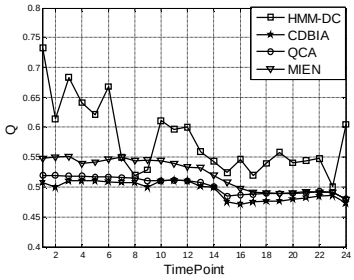

Fig.8 Modularity Q

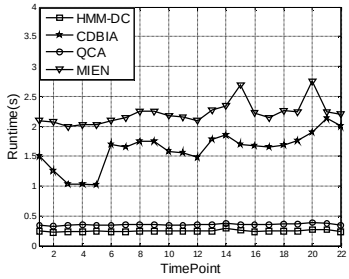

Fig.9 Runtime

Compared with the other algorithms, the HMM_DC algorithm is more stable in community detection and steady in NMI. The performance of the algorithm improved 28% and its modularity values are greater than 0.5.

Facebook dataset: We divide the network by month and typical community dynamic examples discovered in Facebook dataset are shown in Fig. 10-12.

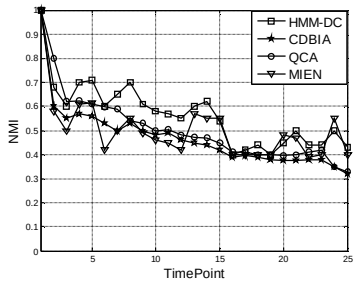

Fig.10 NMI

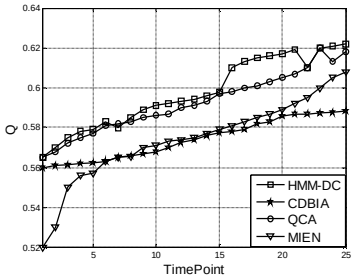

Fig.11 Modularity Q

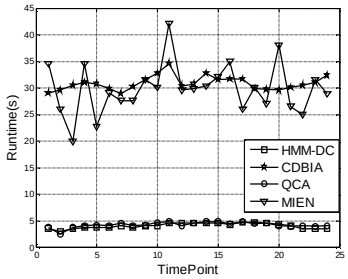

Fig.12 Runtime

Compared with the other algorithms, the performance of HMM_DC algorithm improved 20% and its modularity values are greater than 0.5. In Fig.10-12, we can get that the community structures detected by HMM_DC algorithm are global optimization.

Conclusions

In this paper, the HMM_DC algorithm is proposed based on the Hidden Markov Model to detect the community in dynamic social network. The algorithm transforms the community detection problem to get the optimal status chain in Hidden Markov Model with considering the history information and characteristics in dynamic social network. The algorithm uses the observed chain and status chain to represent the community structure and node information and can identify the community structure without extra information. The results of experiment show that HMM_DC algorithm is available and performs effectively and accurately in identifying the community structure in the dynamic social network and has lower complexity.

From the experiment results, we can get that the HMM_DC algorithm has lower performance in sparse graph and cannot get the global optimization community structure without enough information. Therefore, how to improve the performance of the algorithm in a sparse graph would be considered in our future work.

References

[1] Gregory S. Ordered community structure in networks[J]. Physica A: Statistical Mechanics and its Applications, 2012, 391(8): 2752-2763.

[2] Abrahao B, Soundarajan S, Hopcroft J, et al. On the separability of structural classes of communities[C]. Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2012: 624-632

[3] Lou H, Li S, Zhao Y. Detecting community structure using label propagation with weighted coherent neighborhood propinquity[J]. Physica A Statistical Mechanics and its Applications, 2013, 392: 3095-3105.

[4] Newman M E J. Fast algorithm for detecting community structure in networks[J]. Physical review E, 2004, 69(6): 066133.

[5] Wu Z, Lin Y, Wan H, et al. Efficient overlapping community detection in huge real-world networks[J]. Physica A: Statistical Mechanics and its Applications, 2012, 391(7): 2475-2490.

[6] Symeon Papadopoulos, Yiannis Kompatsiaris, Athena Vakali and Ploutarchos Spyridonos. Community detection in Social Media: Performance and application considerations. Data Min Knowl Disc, Springer, pages 515-554,2012.

[7] Berger-Wolf T Y, Saia J. A framework for analysis of dynamic social networks[C]. Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2006: 523-528.

[8] C. Tantipathananandh, T. Berger-Wolf, and D. Kemne. A framework for community identification in dynamic social networks. KDD, pages 717-726, 2006.

[9] J. Sun, C. Faloutsos, S. Papadimitriou, and P. S. Yu. Graphscope: parameter-free mining of large time-evolving graphs. KDD, pages 687-696, 2007

[10] Liu Y, Liu Q, Qin Z. Community Detecting and Feature Analysis in Real Directed Weighted Social Networks[J]. Journal of Networks, 2013, 8(6): 1432-1439.

[11] Nam P. Nguyen, Thang N. Dinh, Ying Xuan, My T. Thai. Adaptive algorithms for detecting community structure in dynamic social networks. IEEE INFCOM 2011.

[12] Gong M G, Zhang L J, Ma J J, et al. Community Detection in Dynamic Social Networks Based on Multiobjective Immune Algorithm[J]. Journal of Computer Science and Technology, 2012, 27(3): 455-467.

[13] Mitra B, Tabourier L, Roth C. Intrinsically dynamic network communities[J]. Computer Networks, 2012, 56(3): 1041-1053.

[14] Nguyen N P, Dinh T N, Tokala S, et al. Overlapping communities in dynamic networks: their detection and mobile applications[C]. Proceedings of the 17th annual international conference on Mobile computing and networking. ACM, 2011: 85-96.

[15] Huang X, Acero A, Hon H W. Spoken language processing[M]. New Jersey: Prentice Hall PTR, 2001.

[16] Li J, Huang L, Bai T, et al. CDBIA: A dynamic community detection method based on incremental analysis[C]. Systems and Informatics (ICSAI), 2012 International Conference on. IEEE, 2012: 2224-2228..

[17] Dinh T N, Xuan Y, Thai M T. Towards social-aware routing in dynamic communication networks[C]. Performance Computing and Communications Conference (IPCCC), 2009 IEEE 28th International. IEEE, 2009: 161-168.

[18] B. Viswanath, A. Mislove, M. Cha, and K. P. Gummadi. On the evolution of user interaction in facebook. In 2nd ACM SIGCOMM Workshop on Social Networks, 2009.

