计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.190800071

基于衰减系数建立动态蛋白质网络模型进行关键蛋白质预测

戴彩艳　何　菊　胡孔法　丁有伟　李新霞
南京中医药大学人工智能与信息技术学院　南京 210000

摘　要　在生物系统的转变过程中,蛋白质的演化过程并非一成不变,而是动态变化的。通过构造模型的方法来研究蛋白质相互作用网络,可以较好地刻画蛋白质相互作用的演化机制。但是,利用构造模型的方法来研究动态蛋白质相互作用时,应该考虑在蛋白质演化过程中,历史蛋白质随着时间推移对整个演化过程产生作用可能产生的衰减,而不是将不同时刻的蛋白质的作用视为等同或者直接忽略。针对上述情况,提出一种基于衰减系数建立动态蛋白质网络模型的方法。该方法在建立模型的时候采用合理的衰减系数将蛋白质作用的变化情况记录下来,以便于之后研究的开展。通过实验,取合理的衰减系数后,使用相同算法在不同网络模型上运行,结果验证了所提算法的有效性。

关键词:蛋白质相互作用网络;衰减系数;动态蛋白质网络
中图法分类号　TP391

Establishment of Dynamic Protein Network Model Based on Attenuation Coefficient for Key Protein Prediction

DAI Cai-yan, HE Ju, HU Kong-fa, DING You-wei and LI Xin-xia
College of Artificial Intelligence and Information Technology, Nanjing University of Chinese Medicine, Nanjing 210000, China

Abstract　In the transformation process of biological system, the evolution of protein is not static, but dynamic. The evolutionary mechanism of protein interaction can be well described by constructing a model to study protein interaction network. However, when we study protein-protein interaction by using the method of structural model, we should consider the attenuation of historic protein interaction over time in the process of protein evolution, rather than regard the effect of proteins at different times as the same or directly ignore them. In this paper, a method of building dynamic protein network model based on attenuation coefficient was proposed. When establishing the model, a reasonable attenuation coefficient is used to record the changes of protein interaction, which is convenient for later researches. After taking reasonable attenuation coefficient through experiments, using the same algorithm to run on different network models, the results verify the effectiveness of the proposed algorithm.

Keywords　Protein interaction network, Attenuation coefficient, Dynamic protein network

1　引言

随着高通量实验技术[1-2]的发展,出现了种类繁多、规模庞大的蛋白质相关数据。对这些大规模数据抽取构成的蛋白质网络进行研究可以对生命活动的奥秘进行进一步的探索。一些研究者在原始数据库之上通过建立模型的方法对蛋白质网络的变化规律进行研究,以弥补传统研究的缺陷。在建模的过程中,合适的蛋白质网络模型对于研究蛋白质分子间的相互作用规律并在此基础上进行其他研究是十分重要的。

这些年来,研究者们通过不断研究提出了多种不同侧重点的蛋白质相互作用的网络模型。Singh 等[3]通过研究发现,蛋白质网络结构上的无标度分布特性网络的形成与演化之间很可能有着千丝万缕的关联。Sybill 等[4]提出了一个新的蛋白质相互作用演化模型。在该模型中,蛋白质相互作用网络和蛋白质家族相互作用网络同时演化,所形成的两个网络均是无标度网络。Yin 等[5]提出了刻画蛋白质相互作用网络的模型,用数值模拟的方法分析了该模型的演化性质,并将该模型与目前已有的模型进行了比较分析,发现该模型较好地刻画了现实网络的一些性质。Simon 等[6]系统地将需要建立模型的网络分为两类,使用一个时间层次对模型进行综合分析,以评估安全度量的有效性。该模型可以捕获和分析网络系统安全的变化。此外,他们还研究了动态网络发生变化时不同安全度量方法的影响。Harm 等[7]提出了网络可识别性的概念,并将其作为参数化模型集的一个属性,着重研究了网络模型在传递函数方面的区别。该概念确保了在基于测量数据进行识别时,不同网络模型可以彼此区分。Qi 等[8]在将网络中任意顶点的结构特征映射成低维、连续的实值向量的过程中,尽量保留了顶点之间的结构特征关系;并在此基础上提出融合复杂网络结构特征和内容特征的表示学习,以更好地反映出一个网络特征的真实情况,使得学习得到的网络特

基金项目:国家自然科学基金青年科学基金(61906100);江苏省青年基金(BK20180822);江苏省高等学校自然科学研究面上项目(18KJB520040)
This work was supported by the Young Scientists Fund of the National Natural Science Foundation of China (61906100), Jiangsu Province Science Foundation for Youths (BK20180822) and Natural Science Research Projects in Jiangsu Higher Education Institution(18KJB520040).
通信作者:戴彩艳(njucmdai@163.com)

征表示能够有效地被用于各种网络应用中。Luo 等[9]提出一种新的酵母 PPI 网络的局部相互作用密度(LID)概念,基于新的集成策略,将 LID 和蛋白质复合信息结合起来开发出 LIDC 方法,其目的是发现关键蛋白质及其邻域的重要特征,从而提高鉴定效率。Xiao 等[10]提出了在基于动态基因表达构建的动态蛋白质交互网络中识别关键蛋白质的模型。首先,利用时间依赖模型和时间无关模型处理动态基因表达谱。其次,构建了一种新的动态蛋白质交互网络,以增加关键蛋白质的识别数量并提高识别精度。

通过构造模型的方法来研究蛋白质相互作用网,一定程度上克服了基于检测技术获得的蛋白质数据库本身存在的数据采集不准确的缺点,较好地刻画了蛋白质相互作用的演化机制。但是,利用构造模型的方法来研究蛋白质相互作用时,使用的模型要么是静态蛋白质网络,要么是动态蛋白质网络。基于动态蛋白质网络建立相关模型的时候忽略了在蛋白质演化过程中,历史蛋白质作用随着时间推移对整个演化过程产生作用可能衰减的情况。在建立模型的时候如何将蛋白质作用的变化情况保存下来,以便于之后研究的开展,是一个有待探讨的问题。

2 相关工作

静态蛋白质网络是将整个演化过程中的蛋白质看作一个整体,并在此基础上展开研究,比如找其中的关键蛋白质,或者通过相关关联属性预测未知蛋白质的功能等。而动态蛋白质网络更贴近蛋白质的实际演化过程,因为在生物进化中,蛋白质之间的关系是动态变化的,因此能够模拟蛋白质动态演化过程并在此基础上展开研究,得到的结果更符合实际演化的情况。对于蛋白质之间的关系,可以使用网络进行描述,网络中的节点代表不同的蛋白质,而网络中节点之间的边则表明蛋白质之间存在的相互作用。

2.1 动态蛋白质网络时间序列

在对动态蛋白质网络进行建模时,通常是将基因表达数据和大规模的静态蛋白质网络结合起来考虑。可以将 M 个基因在 T 个时间点上的基因表达阵列分为 T 个集合,分别为 $(t_1,t_2,\cdots,t_T)$,T 个时间点经历的时间总和为 t。每个集合表示这 M 个基因在同一时间点的状态,最强科组合成基于时间序列的动态蛋白质网络,如图 1 所示。

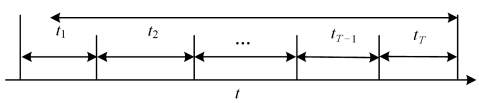


图 1 动态蛋白质网络时间序列示意图

Fig. 1 Time series diagram of dynamic protein network

2.2 蛋白质的演化过程

在不同的时间点上,存在着不同的蛋白质相互作用网络。图 2 显示了简单的蛋白质演化过程。因为真实时间点上蛋白质的种类很多,为了展示蛋白质演化过程,使用 A,B,C 和 D 这 4 个字母来分别代表蛋白质演化过程中在不同时间点上出现的蛋白质。

图 2 简单的蛋白质演化过程

Fig. 2 Simple protein evolution

建立模型的最终目的是方便之后的研究,比如找出蛋白质演化过程中的关键蛋白质,或者是预测在下一时刻蛋白质之间存在的链接关系。这就需要对蛋白质本身的演变过程进行记录。以往记录蛋白质网络存在的链接关系时,是将各个时间点上蛋白质之间出现的链接关系直接记为 1,没有出现链接的关系记为 0。这种方式在时间序列上是不准确的,因为随着时间的推移,历史蛋白质数据所起的作用在衰退,近期产生的蛋白质之间的链接关系所起的作用较大。

动态蛋白质网络体现出随着时间推移整个演化过程的实施。如果其中蛋白质之间的链接关系在演化过程中随着时间推移被认为是前后一样,那么后续基于动态蛋白质模型展开的一系列研究显然会受到影响。因此,如何将演变过程中随着时间推移、比重发生变化的蛋白质链接关系纳入所建模型的范围,以便正确识别动态蛋白质网络中的关键蛋白质并准确预测下一时刻蛋白质的链接关系,值得深入探索。

3 基于衰减系数的动态蛋白质网络的构建

根据基因的表达量确定蛋白质动态属性的基本原则是:如果一个蛋白质对应的编码基因表达量随时间或环境条件波动较大,则该蛋白质是动态的;否则认为其是静态的。因此,可以通过基因表达量变化情况推测某种蛋白质在蛋白质组中的含量随时间的变化情况,从而确定蛋白质处于活动还是非活动状态。该网络构建方法的基本过程如下。

(1)将 M 个基因在 T 个时间点上的基因表达阵列分为 T 个集合,每个集合表示这 M 个基因在同一时间点的状态;利用蛋白质的编码基因在不同时刻的表达值的变化方差将蛋白质划分为动态和静态两类。

假设给定蛋白质网络中有 N 个蛋白质,将基因表达数据分为 T 个部分,以此代表不同时间点的表达量。对于某蛋白质 a,其对应基因在所有时刻的表达值可表示为 E_a。

$$E_a=\{e_{a1},e_{a2},\cdots,e_{aT}\} \tag{1}$$

其中,e_{aT} 表示蛋白质 a 对应基因在 t 时刻的表达量。

蛋白质 a 表达量的方差 σ^2 为:

$$\sigma_a^2=\frac{\sum_{a=1}^{T}(e_{at}-\overline{e_a})^2}{T} \tag{2}$$

其中,$\overline{e_a}$ 为蛋白质对应基因在所有时刻的表达值的平均值,T 表示基因表达数据被分割成的 T 个部分。通过得到的方差可以发现基因在某个生物过程中表达量的变化情况。设置阈值 τ,用来判断蛋白质所处的状态,具体为,

$$IfA(a)=\begin{cases}1, & \sigma_a^2>\tau\\ 0, & \text{else}\end{cases} \tag{3}$$

其中,$IfA(a)$ 值为 1 表示蛋白质 a 为动态蛋白质,为 0 表示蛋白质是静态蛋白质。

(2)根据动态蛋白质以及公开蛋白质网络数据,构建新型动态蛋白质网络。

在构建过程中,随着时间的变化,相同的边在不同时刻出现,在当前时刻来看,其对应的权重计算是不一样的。越早出现的边,它在整个蛋白质进化过程中的作用会随着时间不断发生变化。每个时刻出现的边对应的权重计算方式如下。

对于在 t 时刻出现的蛋白质网络中所含的每一条边 I,其权重 $D(I,t)$ 是一个随时间 t 变化的量,定义为:

$$D(I,t)=\begin{cases}\delta(I,0), & t=0\\ D(I,t-1)\cdot\lambda+\delta(I,t), & \text{otherwise}\end{cases} \quad (4)$$

其中：

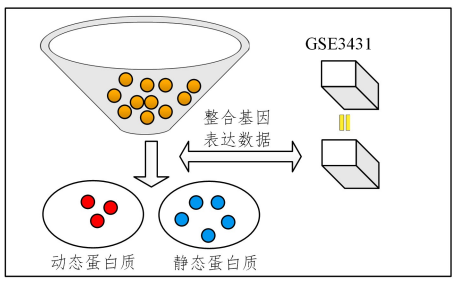
$$\delta(I,t)=\begin{cases}1, & a(t)\text{中含 } I\\ 0, & \text{otherwise}\end{cases}$$

$a(t)$是 t 时刻所有出现的边的集合；$\lambda(0<\lambda<1)$为一个常数，称为衰减系数。

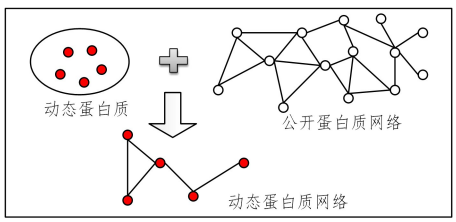
这样建立起来的不同时刻的蛋白质网络中，各条边对应的权重是不同的。

当某一条边随着时间权重不断衰减到一个最小阈值时，可以认为其当前在网络中所起的作用是十分微小的，可以直接将其减去，以节约时间和空间。即 $D(I,t)<\eta$ 时，就可以去掉该边。其中 η 就是针对边的权重设置的最小阈值。

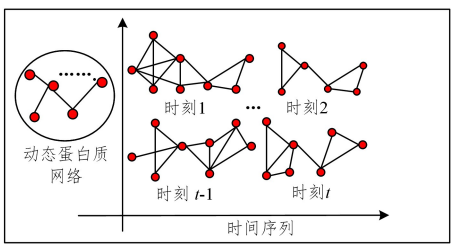
基于衰减系数的动态蛋白质网络的构建过程如图 3 所示。



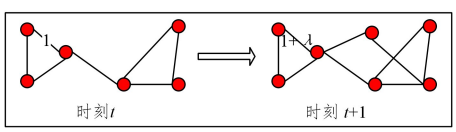
(a)整合基因表达数据，抽取动态蛋白质



(b)结合公开蛋白质网络和动态蛋白质形成动态蛋白质网络



(c)构建不同时刻的动态蛋白质网络



(d)相邻时刻相同边的权重情况

图 3 基于衰减系数的动态蛋白质网络的构建过程

Fig. 3 Construction of dynamic protein network based on attenuation coefficient

因为考虑了历史蛋白质在这个蛋白质演化过程中随时间变化的情况，所以该蛋白质网络模型更为客观，符合生物演化过程。

为了验证基于衰减系数建立的动态蛋白质网络的性能，可以在建立的网络中查找关键蛋白质，并将结果与使用相同方法在其他网络中查找到的结果进行比较。

4 实验结果及分析

4.1 实验环境和条件

实验使用的是 Window 10、i7 处理器、16 GB 内存的计算机，编程语言采用的是 Python 3.8。

4.2 实验数据集

实验中使用的数据有：

(1)基因表达数据 GSE3431[11]，其对应的矩阵有 6 470 行，每一行表示不同基因对应的表达数据；

(2)DIP 中的酵母蛋白质网络[12]，其中包括 5 093 个蛋白质、24 743 条边，部分结构如图 4 所示；

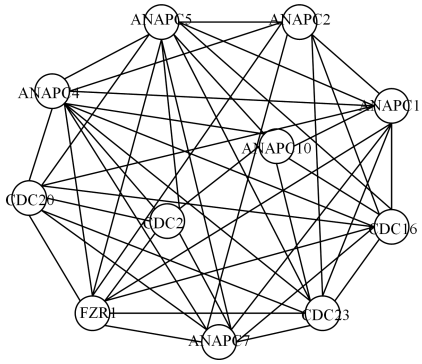


图 4 部分酵母蛋白质网络图

Fig. 4 Part of yeast protein network

(3)集合数据库 MIPS[13]，SGD[14]，DEG[15]和 SGDP[16]得到的 1 285 个关键蛋白质。

4.3 实验结果

4.3.1 衰减系数的实验结果

将动态蛋白质网络分为 36 个时刻，在此基础上对衰减系数取不同值并进行比较。

通过图 5 可以发现，衰减系数在 0.9～0.95 之间时，预测到的关键蛋白质数目最多，由此将衰减系数值设为 0.92。

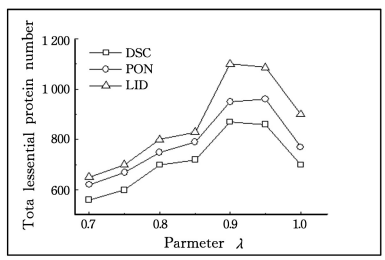


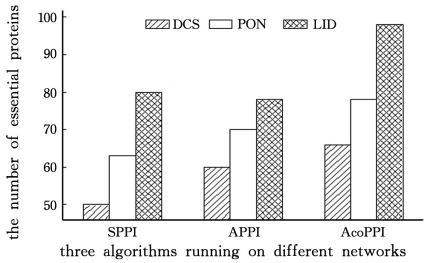
图 5 相同网络上衰减系数的取值对结果的影响

Fig. 5 Effect of attenuation coefficient values in same network on results

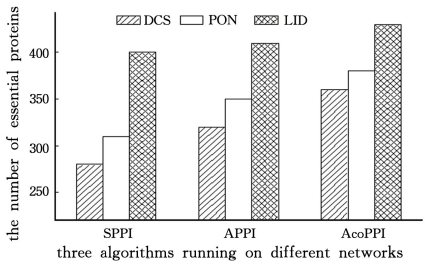
4.3.2 相同算法基于不同网络模型的实验结果

为了验证基于衰减系数建立的动态蛋白质网络的性能，可以在建立的网络中查找关键蛋白质，并将结果与使用相同方法在其他网络中预测到的结果进行比较。

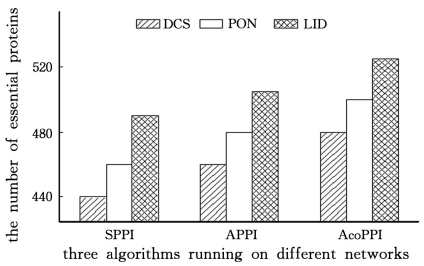
分别在静态网络蛋白质网络(SPPI)、动态蛋白质网络(APPI)[7]以及构建的基于衰减系数的蛋白质网络(AcoPPI)上运行 DCS[17]，PON[18]，LID[9]这 3 类关键蛋白质预测方法，并按照预测出的蛋白质得分进行降序排列。比较前 5%，15%，25%的预测结果中关键蛋白质的数量，结果如图 6 所示。



(a)前 5%预测结果中关键蛋白质的数量



(b)前 15%预测结果中关键蛋白质的数量



(c)前 25%预测结果中关键蛋白质的数量

图 6 前 5%,15%,25%的预测结果中关键蛋白质的数量

Fig. 6 Number of key proteins in the first 5%, 15% and 25% prediction results

通过上面的比较可以发现,基于 AcoPPI 方法建立的模型在考虑了历史蛋白质的影响后,前 5%,15%,25%的预测结果中关键蛋白质数量是最多的。在取 25%的关键蛋白质时,其他两种模型基础上,3 种算法找到的关键蛋白质数目都没有超过 500,而基于 AcoPPI 方法建立的模型查找到的关键蛋白质数目都接近或者超过 500,其中 LID 算法查找到了 518 个关键蛋白质。

下面通过正确率的计算,对基于衰减系数的蛋白质网络模型进行进一步的验证。

正确率的计算公式如下:

$$ACC=\frac{TP+TN}{TP+FN+FP+TN} \tag{5}$$

其中,TP(True Positive)代表本身为正样例的样本被预测为正样例的数量;TN(True Negative)代表本身为负样例的样本被预测为负样例的数量,TP 和 TN 的值表明预测的结果与真实值是一致的,预测结果正确的数量。而 FN(False Negative)代表本身为正样例的样本被预测为负样例的数量;FP(False Positive)代表的是本身为负样例的样本被预测为正样例的数量,FN 和 FP 的值表明预测的结果与真实值是不一致的,预测结果错误的数量。

从图 7 可以看出,基于 AcoPPI 方法建立的模型得到的关键蛋白质准确率是最高的。其原因在于,该模型将蛋白质的历史情况纳入了考量范围。

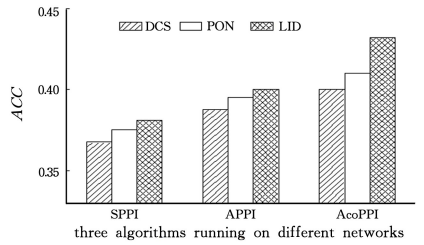


图 7 不同网络模型下关键蛋白质正确率的比较结果

Fig. 7 Accuracy comparison results of key proteins on different network models

结束语 为了考虑蛋白质演化过程中的历史数据对当前数据的影响,提出了基于衰减系数的动态蛋白质网络模型的构造方法。该模型不是简单地将每个时刻蛋白质出现或不出现一概而论,而是根据其对应出现的情况,使用基于衰减系数的 AcoPPI 建模方法记录它们在生物演化过程中所发生的变化,以便于之后研究工作的开展。为了验证其有效性,使用相同的关键蛋白质预测方法在不同的模型上进行了实验。实验结果表明,AcoPPI 建模方法建立的模型能更准确地预测关键蛋白质。下一步的工作是在这个模型的基础上展开进一步的研究,比如在这个模型前提下,如何改进已有的关键蛋白质的查找算法,更为准确地查找关键蛋白质;也可以在此模型上采用卷积神经网络对下一时刻出现的蛋白质之间的关系进行预测。

参 考 文 献

[1] ABBASI W, MINHAS F. Issues in performance evaluation for host-pathogen protein interaction prediction [J]. Journal of Bioinformatics & Computational Biology, 2016, 14(3): i159-i166.

[2] HARM H, PAUL M, AME G. Prediction error identification of linear dynamic networks with rank-reduced noise[J]. Automatica, 2018, 98: 256-268.

[3] SINGH P, SHAKYA M. Comparative evolutionary analysis of cell cycle proteins networks in fission and budding yeast[J]. Cell Biochemistry & Biophysics, 2014, 70 (2): 1167.

[4] SYBILLE D, SEBASTIAN G, CHRISTINE S, et al. Host-pathogen interactions between the human innate immune system and Candida albicans-understanding and modeling defense and evasion strategies[J]. Frontiers in Microbiology, 2015, 6: 625.

[5] YIN R, LI K, ZHANG G, et al. Detecting overlapping protein complexes in dynamic protein-protein interaction networks by developing a fuzzy clustering algorithm[C]// IEEE International Conference on Fuzzy Systems. 2017: 1-6.

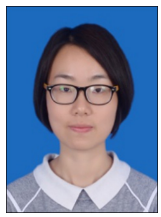
[6] SIMON Y, GE M, JIN B, et al. A systematic evaluation of cybersecurity metrics for dynamic networks[J]. Computer Networks, 2018, 144: 216-229.

[7] HARM H, PAUL M, ARNE G. Identifiability of linear dynamic networks[J]. Automatica, 2018, 89: 247-258.

[8] QI J S, LIANG X, LI Z Y, et al. Representation Learning of Large-Scale Complex Information Network: Concepts, Methods and Challenges[J]. Chinese Journal of Computers, 2018, 41(10): 2394-2420.

[9] LUO J, QI Y. Identification of Essential Proteins Based on a New Combination of Local Interaction Density and Protein Com-

plexes[J]. PLoS One,2015,10(6):e0131418.
[10] XIAO Q,WANG J,PENG X,et al. Identifying essential proteins from active PPI networks constructed with dynamic gene expression[J]. BMC Genomicsvolume,2015,16.
[11] https://www.ncbi.nlm.nih.gov.
[12] http://dip.deo-mbi.ucla.edu/dip/Stat.cgi.
[13] http://mips.helmholtz-muenchen.de/proj/ppi.
[14] https://www.yeastgenome.org.
[15] ZHANG R,LIN Y. DEG 5.0,a database of essential genes in both prokaryotes and eukaryotes[J]. Nucleic Acids Research,2009,37:D455-D458.
[16] http://www-sequence.stanford.edu/ group/ yeast _ deletion_ project/deletions3.html.
[17] PENG W,WANG J,CAI J,et al. Improving protein function prediction using domain and protein complexes in PPI networks [J]. BMC Systems Biology,2014,8(1):35.
[18] LIANG S,ZHENG D,STANGLEY D M,et al. A novel function prediction approach using protein overlap networks[J]. BMC System Biology,2013,7(1):61.



DAI Cai-yan, born in 1985, doctor, lecturer. Her main research interests include bioinformatics and network link prediction.

(上接第 23 页)

[15] KAN J Q,XIE J R,ZHANG H F. Impacts of Social Reinforcement and Edge Weight on the Spreading of Information in Networks[J]. Journal of University of Electronic Science and Technology of China,2014,43(1):21-25.
[16] JENDERS M,KASNECI G,NAUMANN F. Analyzing and predicting viral tweets[C]// Proc. of WWW. 2013:173-182
[17] CHENG J,ADAMIC L,DOW P A,et al. Can cascades be predicted[C]// Proc. of WWW. 2014:42-34.
[18] LERMAN K,GHOSH R. Information contagion:n empirical study of the spread of news on digg and twitter social networks [C]// ICWSM. 2010:54-32.
[19] YANG Y,TANG J,LEUNG C W K,et al. Rain:social role-aware information diffusion[C]// Proc. of AAAI. 2015:34-35.
[20] CHENG J,ADAMIC L,DOW P A,et al. Can cascades be predicted[C]// Proc. of WWW. 2014:43-55.
[21] MUKHERJEE A,VENKATARAMAN V,LIU B,et al. What yelp fake review filter might be doing? [C]// Seventh International AAAI Conference on Weblogs and Social Media. Bellevue,2013.
[22] JING Y P. Reacher of deceptive opinions spam recognition based on deep learning[D]. Shanghai:East China Normal University, 2014.
[23] LI J,OTT M,CARDIE C,et al. Towards a general rule for identifying deceptive opinion spam[C]// Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. 2014:1566-1576.
[24] LAU R Y K,LIAO S Y,KWOK R C W,et al. Text mining and probabilistic language modeling for online review spam detecting [J]. ACM Transactions on Management Information Systems, 2011,2(4):1-30.
[25] OTT M,CHOI Y,CARDIE C,et al. Finding deceptive opinion spam by any stretch of the imagination[C]// Proceedings of the 49th annual meeting of the association for computational linguistics:Human language technologies. Association for Computational Linguistics,Oregon,2011:309-319.
[26] JINDAL N,LIU B,LIM E P. Finding unusual review patterns using unexpected rules[C]// Proceedings of the 19th ACM International Conference on Information and Knowledge Management. ACM,Toronto,2010:1549-1552.
[27] JOULIN A,GRAVE E,BOJANOWSKI P,et al. Bag of Tricks for Efficient Text Classification[J]. arXiv:1607.01759,2016.
[28] SUTSKEVERI,VINYALS O,LE Q V. Sequence to Sequence Learning with Neural Networks[C]// Advances in Neural Information Processing Systems. 2014.
[29] BAHDANAU D,CHO K,BENGIO Y. Neural Machine Translation by Jointly Learning to Align and Translate[J]. arXiv:1409.0473.
[30] BLONDEL V D,GUILLAUME J L,LAMBIOTTE R,et al. Fast unfolding of communities in large networks[J]. arXiv:0803.0476,2008.

CHEN Jin-yin, Ph.D, associate, professor. Her research interests include evolutionary computing, data mining, and deep learning algorithm.