一种基于蛋白质交互网络链接预测的新方法*

李 晶，尚学群，郭 阳，李晓园
（西北工业大学 计算机学院 计算机软件与理论系，西安 710129）

摘 要：当前可用的生物数据在不断地迅速增长，仍有很多生物信息如蛋白质交互信息（protein-protein interaction，PPI）还未被发现，而这些潜在的或未知的信息对生物过程的研究是至关重要的。近年来，对未知生物信息的挖掘和研究吸引了很多人的关注。通过实验检测方法来发现这些信息是非常耗时耗力的，所以链接预测成为一种新的挖掘这些信息的指导方法。基于蛋白质交互网络并融合了基因表达数据信息，从拓扑和基因表达两个方面的信息来构建 PPI 权值网络，提出了一种在权值网络中基于相似度比较的链接预测的新方法来预测 PPI 网络中未知的交互信息。使用 MIPS 数据库评估了实验结果，表明了该算法有很好的准确率和良好的性能。

关键词：蛋白质交互网络；链接预测；权值网络；相关节点集；剪枝
中图分类号：TP391 **文献标志码**：A **文章编号**：1001-3695(2012)11-4060-04
doi:10.3969/j.issn.1001-3695.2012.11.015

New approach of link prediction in PPI network

LI Jing, SHANG Xue-qun, GUO Yang, LI Xiao-yuan
(*Dept. of Computer Software & Theory, School of Computer Science & Engineering, Northwestern Polytechnic University, Xi'an* 710129, *China*)

Abstract: The mount of available biological data is growing at a tremendous pace. However, still a lot of biological information such as some information about protein-protein interaction (PPI) is undiscovered. Those unknown protein information to the study of biological process is essential. Therefore, in recent years, mining and researching unknown biological information has already attracted many people's attention. However, revealing hidden and unknown links of biological network takes the high experimental and time costs. Thus, this paper proposed a new method based on the similarity between proteins to predict implicit or previously unknown links in the weighted PPI networks, and the weight of network could be obtained by combining the topology of PPI network and the inherent information of proteins. It used the MIPS database to evaluate the experimental results show that the algorithm is excellent accuracy performance.

Key words: PPI network; link prediction; weighted network; related nodes set; pruning

0 引言

虽然可以从已经公开的生物数据库中得到很多关于基因组、蛋白质组、蛋白质交互组等生物知识，但遗憾的是，它仅仅只是用来研究细胞过程的小部分，也就是说，凭借现有的生物数据来了解生命的全部过程是远远不够的。因此，发现潜在的、未知的生物信息是生物信息学一项长期的挑战。链接预测就属于研究此问题的一种方法。该方法是由 Kempe 等人[1]提出的，其目的在于通过网络的已知拓扑结构或节点属性等信息来估计在两个尚未连接的节点之间产生一条连边的可能性。无论在生物网络还是社交网络中，对链接预测问题的研究都有着重要的理论和现实意义[2,3]。从挖掘网络中潜在的信息层面来讲，链接预测本质上是对网络演化规律的猜测，高精度的预测算法能够很好地揭示网络的演化行为，有助于理解网络演化的内在机制。链接预测的重要意义还体现在应用方面[4]。对于生物网络中隐而未知链接的揭示是需要耗费高额实验成本的，如果可以预测，而非盲目地检测所有链接，并以此指导实验，就可以节约实验开销。同时，对于演化的在线社会网络而言，可以去预测那些现在尚未结交的用户之间应该是朋友关系，即预测那些未来可能出现但目前并不存在的链接。预测后可以将结果作为朋友推荐信息发送给目标用户，显然这可以帮助用户结交到新的朋友，避免对海量信息进行筛选，也有助于提高相关网站在用户心目中的地位，从而提高用户对该网站的忠诚度。

链接预测在计算机领域主要是提出一些基于马尔可夫链和机器学习过程的算法，但这些方法在物理上不简洁。从网络结构的角度来研究链接预测是一种全新的方式。基于拓扑结构相似性的方法是现在链接预测的主流方法，这种方法有一个重要的假设前提就是如果两个节点之间的相似性越大，那么在它们之间存在链接的可能性也就越大。因此，该方法的一个核心问题就归结为该如何来定义节点之间的相似性[5]。由于这种方法简单可靠，且具有普适性，迄今为止，已经有很多相关的理论和研究方法出现。例如，FOAF[6,7]提出了基于网络的局部特点进行预测，但该方法并不能准确地得到预测结果，因为它只考虑了网络的局部特征，而忽略了捕捉和跟踪网络的全局特征。恰恰相反，RWR[8]算法通过搜索整个网络来发掘潜在

收稿日期：2012-03-15；**修回日期**：2012-05-10 **基金项目**：国家“973”计划资助项目(2012CB316203)

作者简介：李晶(1987-)，女，陕西宝鸡人，硕士研究生，主要研究方向为生物数据挖掘(nwpujing@163.com)；尚学群(1973-)，女，教授，博士，主要研究方向为数据库技术、数据挖掘、生物信息学等；郭阳(1987-)，男，硕士研究生，主要研究方向为数据挖掘、生物信息学；李晓园(1988-)，男，河南新郑人，硕士研究生，主要研究方向为生物数据挖掘.

的信息,该方法的问题在于忽略了网络局部特征的重要性。基于上述两种方法的缺点,Iakovidou 等人[9]将网络的局部和全局特征结合起来,利用多路普聚类来研究链接预测。Popescul 等人[10]利用网络的关联性特征构建了一种结构逻辑回归模型来预测网络中的未知链接。虽然这些方法对链接预测研究都很有效,但是它们在构建权值网络时大多数都只考虑了部分信息。就蛋白质交互网络而言,要么只考虑拓扑信息,要么只考虑蛋白质交互网络内部的交互信息,仅考虑一种信息来预测蛋白质的交互信息是远远不够的。还有一些方法只是简单地在无权网络中进行研究,也就是说,它们都只考虑了网络的拓扑结构,并没有考虑到网络中物体内部自身的特征。由于蛋白质与蛋白质之间的交互主要依赖于生物过程的内部机理,所以用这些方法来预测 PPI 网络是不完整的。基于上述的不足,本文提出了一种把蛋白质自身的内部信息与它们交互网络的拓扑信息结合起来预测 PPI 网络中潜在交互信息的新方法。

本文的贡献可以概括如下:

a)提出了一种构建权值网络的新方法。

在研究过程中,本文把蛋白质自身的内部信息与它们交互网络的拓扑信息结合起来作为网络中两个彼此交互的蛋白质的权重来构建权值网络。

b)提出了一种基于蛋白质之间的相似度来预测潜在交互信息的新方法。

按照本文的定义和方法,在上述构建的权值网络中找出那些可能存在链接关系的点作为候选项,再通过计算这些候选点之间的相似度,真正地确定它们是否有链接关系。

c)做了充分的实验,验证了本文算法具有良好的性能:(a)在不同相似度阈值参数下,验证了预测结果的准确率;(b)在不同完整性水平的蛋白质网络中,对预测结果的平均准确率进行了验证;(c)分别在单一信息构建的权值网络和融合信息构建的权值网络中进行了预测,充分证明了融合信息网络的预测结果更准确。

本文给出了构建权值网络的具体方法,提出 PPI 网络上的有效预测潜在链接的方法及方法中的定义和算法,通过酵母数据来验证本文提出的方法,并对实验结果的准确率进行了分析。

1　权值网络的构建

在本文的方法中,蛋白质交互网络被表示为图 $G=(V,E,W)$。其中:V 表示蛋白质对应的顶点集;E 代表蛋白质相互作用的边集;蛋白质之间的影响也就是权值用 W 记录。为了得到一个很好的权值网络,本文利用网络的拓扑结构信息和蛋白质自身的内部信息这两方面的信息来构建权值蛋白质交互网络。

1.1　拓扑信息权重

如果两个节点拥有更多的共同邻居,那么在它们之间更倾向于存在一条连边。本文基于这个事实,计算两个蛋白质在拓扑结构上的权值。用 W_p 表示来自拓扑结构的权重。为了能准确地描述两点在拓扑网络中的相似度,用网络中两节点拥有的共同邻居数目与两点度的和的比率来计算 $W_p \in [0,1]$。在图 G 中,对于一条边 e,$e=\langle V_i,V_j\rangle \in E$,$W_p(V_i,V_j)$ 按照式(1)可得。式(1)反映出两节点在网络中拓扑结构的交叉程度,即反映了网络中两节点在拓扑结构上的相似程度。

$$w_p(v_i,v_j)=2c/(d(v_i)+d(v_j)) \tag{1}$$

其中:c 代表点 v_i 和 v_j 所拥有的共同邻居节点的数目;$d(v_i)$ 表示点 v_i 的度数;$d(v_j)$ 表示点 v_j 的度数。

1.2　基因表达信息权重

从生物知识中可知,生物体中的蛋白质是通过细胞中基因的转录和翻译以后得到的,也就是说,一个蛋白质本身的功能与它的基因有很大的关系。如果两个蛋白质的基因共表达,那么就认为它们有很大的可能共同协同来完成某个生物功能。所以,对于网络中的两个节点,通过它们的基因表达数据来计算来自内部结构的权重部分,本文用 W_g 表示。通过局部相似度的计算方法和研究微阵列基因表达数据来得到 W_g。局部相似度方法是一种找到两个变化趋势相似的向量的最佳相似度方法。该方法是 Ruan 等人[11,12]在 2006 年提出的,其核心思想是通过计算正值矩阵 $P_{n\times n}$ 和负值矩阵 $N_{n\times n}$ 来得到蛋白质的交互值。通过该方法可知 $W_g \in [-1,1]$。

1.3　融合信息的权重

得到来自两个不同信息的权重值之后,本文把这两方面的权重结合起来定义网络中一对蛋白质之间真正的权重值。具体的计算方法如式(2)所示:

$$w=\alpha\times w_g+\beta\times w_p \tag{2}$$

其中:α 和 β 是事先定义好的参数。通过生物知识可知,生物体之间的交互主要来自于生物体内部机理,所以对于网络中两点权重的计算,W_g 比 W_p 重要得多,因而 α 要远远大于 β。

2　链接预测方法的描述

针对生物网络的特点,在 PPI 权重网络中提出了一种基于蛋白质相似度的方法来预测蛋白质与蛋白质之间潜在的交互信息。该方法大体概括为以下几个步骤:

a)对于网络中的一点 P,先找出它的相关节点集,也就是它的所有邻居节点构成的集合。

b)遍历网络中除了点 P 之外的所有点,如果该点与点 P 的相关节点集中的所有节点都有边相连,那么就把该点作为点 P 的相关节点集的候选项。

c)通过计算任意两个候选项之间的相似度来确定这两点之间是否有链接关系。

2.1　产生每个种子节点的候选项

定义 1　一个点的相关节点集。在图 G 中,对于 V 中的一个节点 P,定义由网络中点 P 的所有邻居节点组成的集合为点 P 的相关节点集。为了方便,用 S_p 来表示点 P 的相关节点集。

定义 2　一个相关节点集的所有候选项。在图 G 中,对于 V 中的一个节点 P,找到它的相关节点集合 S_p 后,定义网络中与 S_p 所有节点都有边的点为 S_p 的候选项。

下面通过一个例子来解释上述的定义。如图 1 所示,当考虑 V_a 为种子节点时,图 1 中由节点 V_a 的所有邻居节点 V_1、V_2、V_3、V_4 组成的集合作为 V_a 的相关节点集 S_a;然后遍历整个网络,发现点 V_a、V_b、V_c 与 S_a 中的所有节点 V_1、V_2、V_3、V_4 都有边相连,所以点 V_a、V_b、V_c 就是 S_a 的候选项。当考虑 V_b 为种子节点时,图 1 中由节点 V_b 的所有邻居节点 V_1、V_2、V_3、V_4、V_5 组成的集合作为 V_b 的相关节点集 S_b;最后遍历整个网络,发现点 V_b、V_c 与 S_b 中的所有节点 V_1、V_2、V_3、V_4、V_5 都有边连接,所以点 V_b、V_c 就是 S_b 的候选项。

从上述的例子中可以发现,一个相关节点集候选项节点都拥有很多的共同邻居节点。基于如果两点节点拥有很多共同邻居节点,那么这两个节点之间很可能存在链接关系的事实,猜想某个相关节点集的候选项之间也可能存在链接关系。所以本文通过计算这些候选项之间的相似度来确定它们之间是否真正存在链接关系。

2.2 计算候选项之间的相似度

2.2.1 网络中一个点的向量表示

通常在有权网络中,两点之间的相似度是通过它们的向量计算的。因此,为了得到一个相关节点集候选项之间的相似度,就必须知道它们的向量表示。

定义 3 候选项的向量表示。在图 G 中,一个节点 P 的相关节点集 S_p,对于 S_p 中一个候选项节点(R),定义 R 的向量是由 R 与 S_p 中所有节点的权值构成。

当 V_a 作为种子节点考虑时,如图 1 所示,S_a 中有四个节点,分别为(V_1,V_2,V_3,V_4),S_a 有三个候选项 V_a、V_b、V_c,所以 V_a 的向量记做 $a=\langle a_1,a_2,a_3,a_4\rangle$。其中 a_1、a_2、a_3、a_4 分别为点 V_a 与 S_a 中 V_1、V_2、V_3、V_4 边上的权重值。V_b 的向量为 $b=\langle b_1,b_2,b_3,b_4\rangle$。其中 b_1、b_2、b_3、b_4 分别为点 V_b 与 S_a 中 V_1、V_2、V_3、V_4 边上的权重值。特别值得注意的是,对一个相关节点集的所有候选项进行向量表示时,向量中的元素必须一一对应。也就是说,a_1 与 b_1 分别是 V_a、V_b 与 S_p 中同一个 V_1 的权重值。当 V_b 作为种子节点考虑时,如图 1 所示,S_b 中有五个节点,分别为(V_1,V_2,V_3,V_4,V_5),S_b 有两个候选项 V_b、V_c。根据定义,这时 V_b 的向量表示为 $b=\langle b_1,b_2,b_3,b_4,b_5\rangle$。其中 b_1、b_2、b_3、b_4、b_5 分别为点 V_b 与 S_b 中 V_1、V_2、V_3、V_4、V_5 边上的权重值。由此可见,在本文的方法中,网络中一个点的向量表示并不是一成不变的,而是随着种子节点的相关节点集中节点个数不断变化的。

2.2.2 网络中两个节点相似度的计算

目前,无论是社交网络还是生物网络,都已经提出了很多计算两个向量相似度的方法,最常见的就是用欧式距离和 cosine 来计算。但这两种方法都有很大的缺点:只是简单地比较两个向量中的元素,因而这两种方法只能反映两个向量之间的距离,而不能反映出它们的变化趋势的相似度。鉴于上述缺点,本文采用相关系数(pearson product-moment correlation coefficient,PCC)来计算两个向量之间的相似度。PCC 是 Pearson 提出的,由于它能够很好地反映出两个变量变化趋势的优点而被广泛地用来衡量两个变量的相似度。PCC 的计算如式(3)所示:

$$r(v_i,v_j)=\frac{\sum xy-\frac{\sum x\sum y}{n}}{\sqrt{\left(\sum x^2-\frac{(\sum x)^2}{n}\right)\left(\sum y^2-\frac{(\sum y)^2}{n}\right)}} \tag{3}$$

其中:r 代表两个相似度的计算结果,X 代表向量 V_i 中的元素;Y 代表向量 V_j 中的元素。因为向量 V_i 与向量 V_j 的维数相等,本文用 n 表示。

计算完两个点的相似度后,如果计算结果大于预先设定的相似度阈值 ε,就认为这两点之间存在链接。对应于 PPI 网络中,就认为这两个蛋白质存在交互关系。

2.3 算法中的剪纸策略

通过上文提出的预测方法,可以进行有效的预测,但无论是在选择种子节点还是在预测结果中都存在很大的重复。为了去除上述的重复,提高算法的效率,本文提出了三个剪枝策略对该方法进行修改和完善。

2.3.1 剪枝 1

众所周知,生物数据存在很大的噪声,反映到 PPI 网络中就是,网络中存在一些点很少与其他点交互,甚至还有一些孤立点。显然,根据本文的方法把这些点选为种子节点来进行预测是没有意义的,因为它们的相关节点很少甚至没有,所以在算法中将忽略这些点。在后面实验中,只选择度大于 3 的节点作为种子节点。这种有目的的筛选在不降低预测结果的同时也提高了算法的效率。

2.3.2 剪枝 2

PPI 网络是一个十分复杂的网络,有可能存在一些点它们有着相同邻居节点,根据定义,也就是这些点有相同的节点相关集。如图 1 所示,点 V_b 和 V_c 有相同的相关节点集 V_1、V_2、V_3、V_4、V_5。对于同一个节点相关集而言,其候选项集也是相同的。所以无论考虑点 V_b 还是点 V_c,节点相关集 V_1、V_2、V_3、V_4、V_5 的候选项都是 V_b 和 V_c。也就是说,如果不进行剪枝,那么 V_b 与 V_c 的相似度将会按照相同的向量计算两边。在大型的 PPI 网络中,可能存在很多点它们都有着相同的节点相关集,这样就是对一对节点按照相同的数据多次重复计算它们的相似度,这显然带来了内存和时间上的浪费。为了避免这一问题,本文提出了剪枝策略,对具有相同节点相关集的节点,只选取其中一个点作为种子节点,其他节点都将被忽略。这样对于庞大的复杂网络,大大提高了算法的效率。

2.3.3 剪枝 3

上述两种剪枝策略已经使算法很有效,但本文的方法中仍存在问题。如图 1 所示,当考虑 V_a 为种子节点时,S_a 中有 V_1、V_2、V_3、V_4 四个节点,S_a 的候选项有 V_a、V_b、V_c。这时利用 4 维向量计算 V_b 与 V_c 的相似度。而当 V_b 是种子节点时,S_b 中有 V_1、V_2、V_3、V_4、V_5 五个节点,S_b 的候选项有 V_b、V_c,这时必须利用五维向量计算 V_b 与 V_c 的相似度。不难看出,这时对于同一对节点有两个不同的相似度。所以,对于同一对节点取它们相似度的最大值来避免这种重复。

2.4 算法描述

输入:图 G 中点的集合 V,图 G 中边的集合 E,相似度阈值 ε,用于存放一个点的节点相关集 V',用于存放一个节点相关集的候选项集 Q。

输出:所有预测的链接:E'。

初始化:$V'=\varnothing$,$E'=\varnothing$,$Q=\varnothing$

过程:GENERATE (V,E,V',Q,ε)

```
1 for each node P in V
2 找到点 P 的相关节点集 Sp;
3 if degree(Sp) >3,将 Sp 中所有点存放到 V'中
4 for each node q in V except p
  //对 V 中除了点 p 的所有节点 q 都检验
5 if 点 q 与 V'中的所有点都相连
6 Q←q; //q 就为相关节点集 Sp 的候选项
7 endif;
8 endfor;
9 else 转到 1
10 endif;
11 while (Q ≠∅)
     Q← P ; // 将点 P 放到 Q 中
     calculatePCC(Vi,Vj); // 计算 Q 中任意两点 Vi、Vj 的相似度
12 if r(Vi,Vj) > ε
   then E' putNode(Vi,Vj, r(Vi,Vj));
   //如果相似度大于相似度阈值,这两点就作为预测的结果
13 endif
```

14 检验 Q 中的除点 P 之外的所有点 S,如果 $S_p == S_s$,从 V 中删除 S 点
15 endif
16 endwhile
17 V′ = ∅, Q = ∅;
18 endfor
19 找出 E′中所有重复的边,只保留重复项中相似度最大的一条边
20 printf(E′);
end

因为本文方法是基于网络拓扑结构进行预测的,所以考虑孤立点和邻居节点很少的点没有意义,为了尽快有效地预测潜在的信息,本文不考虑网络中度小于 3 的节点。当且仅当顶点 p 的相关节点集 $S_p > 3$ 时,将 S_p 中的所有节点加入到 V'中(第 3 行),搜索全图中的除点 P 之外的所有节点。如果有点与 S_p 中的所有节点都相连,将该点加入到队列 Q 中,作为 S_p 的候选项(第 4、5、6 行)。计算候选项中任意两点之间的相似度,如果大于相似度阈值,就预测这两点之间存在链接关系(第 11、12 行)。检验所有候选项的节点相关集是否与 S_p 相同,如果相同,从图 V 中删除该点(第 14 行)。将 V'、Q 清空,按照上述方法再考虑图中其他节点,直到图中所有的节点都被考虑过为止。检查输出结果集,删除重复边,只保留两点中权重最大的边(第 16 行)。

根据算法,先找出 PPI 网络中的每个种子节点相关的节点集,然后找到设置此相关节点的所有候选项,最后通过上述定义的相似性,预测这些候选项之间的链接关系。显然,本文算法的复杂度为 $O(n^2)$。

3　实验

3.1　数据集描述和分析

实验以酵母蛋白质相互作用网络作为研究对象,因为酵母是所有物种中蛋白质相互作用数据最为完备的,以 1998 年 Cho 等人发布的酵母基因表达数据作为微阵列数据。这些数据均可以从斯坦福基因数据库和其他生物数据库中得到。它们是通过观察酵母细胞在 α 条件下从分裂前期的 G_1 期到分裂后期的 M 期表达收集的。

3.2　结果对比

3.2.1　相似度阈值变化对实验结果的影响

实验 1　以 MIPS 数据库中的酵母蛋白质交互网络作为标准(http://mips.helmholtz-muenchen.de/),随机从网络中剔除 500 条边,构成一个不完整的网络。在这个不完整网络中按照本文的算法进行预测,再用预测后的结果与原有的标准网络进行比较。考察在阈值 ε 变化的情况下,实验结果在精确度上的变化。其中,准确率 P 按式(4)计算可得

$$p = \frac{n}{m} \tag{4}$$

其中:n 为在预测结果在标准网络中存在的链接数,m 为预测的总链接数。图 2 为实验结果精确度随 ε 的变化。

从图 2 中可以看出,随着相似度阈值 ε 的增大,预测结果的准确率也在不断增大。当 $\varepsilon = 0.9$ 时,预测结果的最高准确率可达 45%;当 $\varepsilon = 0.4$ 时,预测结果的准确率也能达到 27%,平均准确率为 32%。显然,本文的算法在链接预测中有相对较好的准确率。

3.2.2　网络不同完整性水平对实验结果的影响

实验 2　通过不同完整性水平上的网络来观察实验结果准确率的变化。首先从原酵母蛋白质交互网络中随机剔除一些边,分别保留 40%、50%、60%、70%、80%、90% 条边,形成不同的不完整网络。为了观察网络的整体水平,分别取 $\varepsilon = 0.9$, 0.8, 0.7, 0.6, 0.5, 0.4 在每个网络上进行预测,得到的结果按照式(4)分别计算出不同相似度阈值上的准确率,然后对每个网络的 6 个不同相似度 ε 上的准确率进行平均,得到每个网络上的平均准确率,再对每个网络上的平均准确率进行比较分析。图 3 为实验结果随观察点的变化。

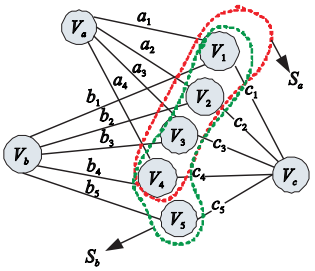

图1　例子

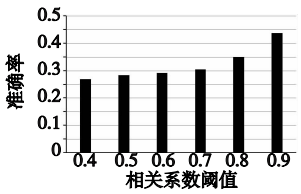

图2　实验结果精确度随ε的变化

从图 3 中可以看出,观察点越多实验结果的准确率就越大。当网络中有 90% 的边是已知时,预测结果的最高平均准确率可达 47%;当只有 40% 的边是已知时,预测结果的平均准确率也能达到 32%。从上述的分析结果可以看出,本文的算法能够进行有效地预测。

3.2.3　权重信息对实验结果的影响

实验 3　通过不同权重信息构成的权值网络来观察不同完整性水平下网络预测结果准确率的变化。首先运用只有拓扑权重信息构成一个蛋白质交互网络,利用本文的相似度预测方法进行链接预测并对准确率进行评价;然后运用只有基因表达数据权重信息构成一个蛋白质交互网络,利用本文的相似度预测方法进行链接预测并对准确率进行评价;最后将两种权重信息融合在一起构成一个蛋白质交互网络。从上述三种原始酵母蛋白质交互网络中随机地剔除一些边,分别保留 40%、50%、60%、70%、80%、90% 条边,形成不同的不完整网络。运用本文基于相似度的方法分别在相关系数大于 0.4 的情况下观察不同权重下预测结果的准确率。图 4 为实验结果比对。

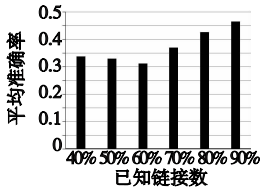

图3　实验结果随观察点的变化

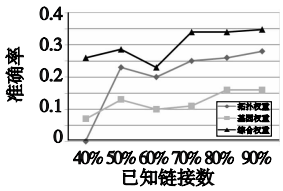

图4　实验结果比对

从图 4 中可以看出,综合权重网络的整体准确率优于单个权重网络的准确率。当网络中有 90% 的边是已知时,综合权重网络准确率可达 35%,而基因表达数据构成的权重网络的准确率只能达到 16%,拓扑权重网络的准确率只能达到 27%。从上述的分析结果可以看出,单一信息在链接预测的局限性和将两种权重信息融合起来进行链接预测可以达到很好的效果。

4　结束语

本文提出了一种基于蛋白质之间相似度的方法来预测 PPI 网络中潜在的、隐藏的链接。本文将 PPI 网络看做是一个有权图,根据网络中两节点的拓扑结构和权重信息,通过计算它们的相似度来预测它们是否存在链接关系,而且通过真实的酵母蛋白质交互网络观察预测结果的准确

(下转第 4078 页)

个未分配的远程任务分配执行,解决了 DS 算法中存在的无本地数据节点闲置问题。

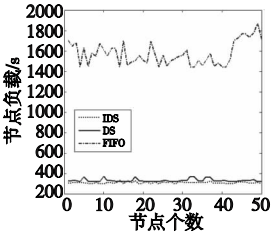

图13 10 G bit环境下节点负载的比较(三种算法)

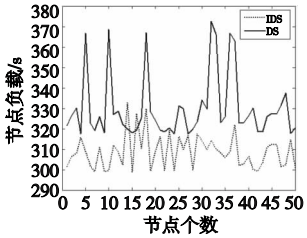

图14 10 G bit环境下节点负载的比较(两种算法)

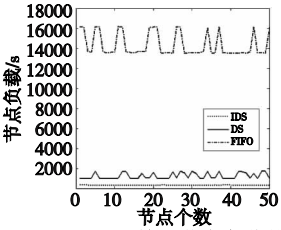

图15 1 G bit环境下节点负载的比较(三种算法)

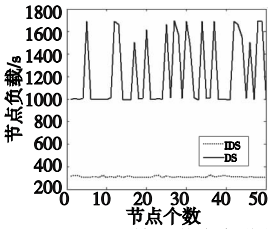

图16 1 G bit环境下节点负载的比较(两种算法)

4 结束语

本文提出了一种基于动态等待时间阈值的延迟调度算法。该算法在等待时间阈值及节点闲置问题上对原有的算法进行改进,对每个作业设置一个动态的等待时间阈值,以调整作业的数据本地性与公平性;对无本地数据节点增设一个最大等待时间,较好地解决了节点闲置问题,使其能够适应节点不会很快空闲的情况。仿真实验结果表明,改进的延迟调度算法通过以上两方面的改进,在作业完成时间及负载均衡方面优于已有延迟调度算法的性能。

参考文献:

[1] DEAN J, GHEMAWAT S. MapReduce: simplified data processing on large clusters[J]. Communication ACM, 2008, 51(1): 107-113.

[2] ISARD M, BUDIU M, YU Yuan, *et al*. Dryad: distributed data-parallel programs from sequential building blocks[C]//Proc of ACM SIGOPS/EuroSys European Conference on Computer Systems. New York: ACM Press, 2007: 59-72.

[3] Hadoop [EB/OL]. (2011-12-18) [2012-03-12]. http://hadoop.apache.org.

[4] WANG Guo-hui, NG T S E. The impact of virtualization on network performance of amazon EC2 data center[C]//Proc of the 29th Conference on Information Communications. Piscataway, NJ: IEEE Press, 2010: 1163-1171.

[5] FISCHER M J, SU Xue-yuan, YIN Yi-tong. Assigning tasks for efficiency in Hadoop: extended abstract[C]//Proc of the 22nd ACM Symposium on Parallelism in Algorithms and Architectures. New York: ACM Press, 2010: 30-39.

[6] ZAHARIA M, BORTHAKUR D, SARMA J S, *et al*. Delay scheduling: a simple technique for achieving locality and fairness in cluster scheduling[C]//Proc of European Conference on Computer Systems. Paris: [s. n.], 2010: 265-278.

[7] GHEMAWAT S, GOBIOFF H, LEUNG S T. The Google file system [C]//Proc of the 19th ACM Symposium on Operating Systems Principles. New York: ACM Press, 2003: 29-43.

[8] WHITE T. Hadoop: the definitive guide[M]. [S. l.]: O'Reilly Media, Inc, 2009.

[9] JIN Jia-hui, LUO Jun-zhou, SONG Ai-bo, *et al*. BAR: an efficient data locality driven task scheduling algorithm for cloud computing [C]//Proc of the 11th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing. Washington DC: IEEE Computer Society, 2011: 295-304.

[10] 金嘉晖,罗军舟,宋爱波,等. 基于数据中心负载分析的自适应延迟调度算法[J]. 通信学报, 2011, 32(7): 47-56.

[11] Max-min fairness[EB/OL]. (2011-11-16)[2012-03-12]. http://en.wikipedia.org/wiki/Max-min_fairness.

[12] CHANG F, DEAN J, GHEMAWAT S, *et al*. Bigtable: a distributed storage system for structured data[J]. ACM Trans on Comput Systems, 2008, 26(2): 1-26.

(上接第 4063 页)率,并对本文算法的有效性进行了分析。但是本文仍有不足之处,如针对该方法只在一个数据集上进行了实验验证。在多个数据集上验证本文的算法和将该方法拓展到动态的 PPI 网络进行链接预测是笔者进一步研究的任务。

参考文献:

[1] KEMPE D, KLEINBERG J M, TARDOS E. Maximizing the spread of influence through a social network[C]//Proc of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2003: 137-146.

[2] ZHOU Tao, LV Lin-yuan, ZHANG Yi-cheng. Predicting missing links via local information[J]. The European Physical Journal B, 2009, 71(4): 623-630.

[3] 吕琳媛. 复杂网络链接预测[J]. 电子科技大学学报, 2010, 39(5): 651-661.

[4] 王琳,商超. 无标度网络中的链路预测问题研究[J]. 计算机工程, 2012, 38(3): 67-71.

[5] LIBEN-NOWELL D, KLEINBERG J. The link prediction problem for social networks[C]//Proc of International Conference on Information and Knowledge Management. 2003: 556-559.

[6] CHEN Ji-lin, GEYER W, DUGAN C, *et al*. Make new friends, but keep the old recommending people on social networking sites[C]//Proc of the 27th International Conference on Human Factors in Computing Systems. 2009: 201-210.

[7] MAXWELL J C. A treatise on electricity and magnetism[M]. 3rd ed. Oxford: Clarendon, 1892: 68-73.

[8] PAN Jia-yu, YANG H I, FALOUTSOS C, *et al*. Automatic multimedia cross-modal correlation discovery[C]//Proc of the 10th ACM-SIGKDD International Conference on Knowledge Discovery and Data Mining. 2004: 653-658.

[9] IAKOVIDOU N, SYMEONIDIS P, MANOLOPOULOS Y. Multiway spectral clustering link prediction in protein-protein interaction networks[C]//Proc of the 10th IEEE International Conference on Information Technology and Applications in Biomedicine. 2010: 1-4.

[10] POPESCUL A, UNGAR L H. Statistical relational learning for link prediction[C]//Proc of at IJCAI Workshop on Learning Statistical Models from Relational Data. 2003.

[11] RUAN Quan-song, DUTTA D, SCHWALBACH M S, *et al*. Local similarity analysis reveals unique associations among marine bacterioplankton species and environmental factors[J]. Bioinformatics, 2006, 22(20): 2532-2538.

[12] RUAN Quan-song, STEELE J A, SCHWALBACH M S, *et al*. A dynamic programming algorithm for binning microbial community profiles [J]. Bioinformatics, 2006, 22(12): 1508-1514.