```
5:    for each g∈G′ do
6:      if g is not in N then
7:        create a node containing g in N;
8:      endif
9:      add an edge connecting g and d to N;
10:   end for
11:   P′←all the phenotypes of d;   //疾病 d 的所有表型
12:   for each p∈P′ do
13:     if p is not in N then
14:       create a node containing p in N;
15:     endif
16:     add an edge connecting p and d to N;
17:   end for
18: end for
19: return N;
```

算法 1 的基本思想为:对每一种疾病,遍历其每一个致病基因和每一个表型,如果是第 1 次出现,则在疾病信息网络中建立一个相应类型的节点以及表示疾病与基因关系(*GD*)、疾病与表型关系(*DP*)的边.由于在疾病信息网络中,基因节点和表型节点之间通过疾病节点连接,而表型集合远大于基因集合(|*P*|>>|*G*|),因此,算法 1 的时间复杂度为 $O(|D|\times|P|)$.

我们利用图数据库 Neo4j[26]存储疾病信息网络.一方面,Neo4j 支持兼容 ACID 特性的事务操作,可以对疾病信息网络进行基本的查询;另一方面,对于高度关联的数据查询速度更快且提供了一个可视化查询平台,便于直观地查看实体与实体间联系组成的结构.

4 表型相似基因搜索

4.1 gSim-Miner算法框架

我们设计了 gSim-Miner 算法实现在疾病信息网络中搜索与查询基因具有最大表型相似度的 *k* 个基因.考虑实际疾病信息网络的稀疏性,相似的基因可能并不共享直接邻居,而且较长的疾病路径会降低相似性搜索的临床意义.因此,本文仅定义 GDPDG 作为基因相似性的疾病路径,这在一定程度上也降低了计算基因连通路径的计算代价.gSim-Miner 算法的基本思想在于遍历每个候选基因,计算候选基因与查询基因间的表型相似度寻找挖掘结果.为此,需要分别计算从查询基因出发满足元路径 GDPDG 返回自身和连通候选基因的路径实例数,以及从候选基因出发满足元路径 GDPDG 返回自身的路径实例数.

算法 2 描述了 gSim-Miner 算法的基本框架.其中,步骤 2、步骤 5 和步骤 6 中均涉及到 *Ins* 函数的计算.*Ins* 函数是通过遍历第 3 节构建的疾病信息网络 *N* 得到的.根据已有的疾病信息网络,我们可以得到给定查询基因 *g* 以及候选基因 *g′*的表型集合,进而可以计算出|*Ins*(*gDPDg*)|和|*Ins*(*g′DPDg′*)|.接着,通过比较两个基因表型集合中的元素,从而计算出基因 *g* 和基因 *g′*通过相同表型连通的路径实例数,即|*Ins*(*gDPDg′*)|.分析算法 2 可知,Naïve 算法时间复杂度为 $O(np)$,其中,*n* 表示候选基因集合的个数,*p* 表示基因对应表型数量的平均值.

算法 2. gSim-Miner 算法框架(Naïve 算法).

输入:疾病信息网络 *N*,参数 *k*,查询基因 *g*;

输出: top-*k* 表型相似基因 *R*.

```
1:  R←{g};
2:  count the number of path instances of gDPDg;   //compute |Ins(gDPDg)|
3:  G←all genes in N;
```

4: **for** each $g'\in G\backslash g$ **do**
5: count the number of path instances of $g'DPDg'$; //compute $|Ins(g'DPDg')|$;
6: count the number of path instances of $gDPDg'$; //compute $|Ins(gDPDg')|$;
7: **if** $|R|<k$ **then**
8: add g' to R;
9: **elseif** $Sim(g,g')\geqslant \mathrm{MIN}\{Sim(g,g'')|g''\in R\}$ **then**
10: update R with g';
11: **end if**
12: **end for**
13: **return** R;

请注意:对于查询基因,其必与自身表型相似,因此可以直接将其加入到搜索结果中(见算法 2 步骤 1).在特定的应用情景下,也可以将其排除(步骤 1 初始化 R 为空集).

4.2 gSim-Miner算法

容易看出,算法 2 执行效率较低.原因在于:(1) 计算路径实例数涉及大量关于图连通性测试的计算;(2) 候选基因为网络中所有基因节点.针对问题(1),为提高计算基因表型相似度计算效率,我们物化了疾病路径 *GDP*,并以字典{⟨gene,phenotype⟩,weight}的形式存储,其中,weight 是通过计算所有符合 *GD* 与 *DP* 路径实例数乘积的总和得到.这样,如果基因间存在相同表型,那么基因间便可以连通,能够快速地根据基因查找到对应的表型集合,并计算出基因间的表型相似度.针对问题(2),我们基于如下观察设计了候选基因的剪枝策略:

定理 1. 在疾病信息网络中,对任意基因 g,令 $P(g)$表示所有与 g 通过 *GDP* 路径实例连通的表型节点集合,即 $P(g)=\{p\in P|\exists GDP$ 路径实例连通 p 和 $g\}$.如果 $P(g)\cap P(g')=\varnothing$,那么基因 g 与基因 g'必定不相似.

证明:若基因 g 与基因 g'相似,那么必然存在一条疾病路径实例通过表型节点 p 连通基因 g 与基因 g';反之,若 $P(g)\cap P(g')=\varnothing$,即不存在任意一条疾病路径实例使得基因 g 与基因 g'关联,那么二者必定不相似. □

根据定理 1,我们设计剪枝策略 1.

剪枝策略 1. 令 g 为查询基因,g'为候选基因.若 $P(g)\cap P(g')=\varnothing$,那么减去 g'.

给定基因 g,令$|P(g)|$表示所有与 g 通过 *GDP* 路径实例连通的表型节点数,令$|Ins(gDPDg)|=\sum_{i=1}^{|P(g)|}p_i^2$ 表示基因 g 满足元路径 *GDPDG* 到自身的路径实例数.遍历存储 *GDP* 的字典,对候选基因到自身的路径实例数$|Ins(g'DPDg')|$进行升序排列,令 $\mathrm{MAX}\{|Ins(g'DPDg')||g'\in G\}$表示候选基因 g'排完序后的最大值.由定理 1,我们可以得到如下性质:

性质 1. $|Ins(gDPDg')|\leqslant \mathrm{MIN}(|Ins(gDPDg)|,\mathrm{MAX}\{|Ins(g'DPDg')||g'\in G\})$.

根据性质 1,我们设计剪枝策略 2.

剪枝策略 2. 令 g 为查询基因,g'为候选基因,当前已知 top-k 表型相似基因的表型相似度最小值为 s.若 $2\times \mathrm{MIN}(|Ins(gDPDg)|,\mathrm{MAX}\{|Ins(g'DPDg')||g'\in G\})/(|Ins(gDPDg)|+|Ins(g'DPDg')|)<s$,那么剪去 g'.

证明:给定查询基因 g,$\forall g'(g'\in G$ 且 $g'\neq g)$,$|Ins(gDPDg')|=|Ins(gDPDg)\cap Ins(g'DPDg')|$,那么,

$|Ins(gDPDg')|\leqslant \mathrm{MIN}(|Ins(gDPDg)|,|Ins(g'DPDg')|)\leqslant \mathrm{MIN}(|Ins(gDPDg)|,\mathrm{MAX}\{|Ins(g'DPDg')||g'\in G\})$.

因此,

$$Sim(g,g')=2\times|Ins(gDPDg')|/(|Ins(gDPDg)|+|Ins(g'DPDg')|)\leqslant$$
$$2\times \mathrm{MIN}(|Ins(gDPDg)|,\mathrm{MAX}\{|Ins(g'DPDg')||g'\in G\})/(|Ins(gDPDg)|+|Ins(g'DPDg')|).$$ □

算法 3 给出了带剪枝策略的 gSim-Miner 算法.

算法 3. gSim-Miner 算法.

输入:疾病信息网络 N,参数 k,查询基因 g;

输出:top-k 表型相似基因 R.

1: R←{g};
2: count the number of path instances of $gDPDg$; //compute |Ins($gDPDg$)|
3: P(g)←all phenotypes of g satisfying disease path GDP;
4: G←all genes in N;
5: **for** each $g'\in G\backslash g$ **do**
6: P(g')←all phenotypes of g' satisfying disease path GDP;
7: **if** $P(g)\cap P(g')=\varnothing$ **then**
8: remove g' from G and **continue**; //剪枝策略 1
9: **end if**
10: count the number of path instances of $g'DPDg'$; //compute |Ins($g'DPDg'$)|
11: **end for**
12: G←sort g' in ascending orderaccording to thenumber of path instances
13: **for** each $g'\in G\backslash g$ **do**
14: count the number of path instances of $gDPDg'$; //compute |Ins($gDPDg'$)|
15: **if** $2\times \mathrm{MIN}(|Ins(gDPDg)|,\mathrm{MAX}\{|Ins(g'DPDg')||g'\in G\})/(|Ins(gDPDg)|+|Ins(g'DPDg')|)<\mathrm{MIN}\{Sim(g,g'')|g''\in R\}$ **then**
16: remove g' and **goto** Step 13; //剪枝策略 2
17: **end if**
18: **if** $|R|<k$ **then**
19: add g' to R
20: **elseif** $Sim(g,g')\geq \mathrm{MIN}\{Sim(g,g'')|g''\in R\}$ **then**
21: update R with g';
22: **end if**
23: **end for**
24: **return** R;

5 实 验

5.1 实验环境

本文利用真实数据集验证本文提出的疾病信息网络的适用性及表型相似基因搜索算法的有效性、执行效率和可扩展性.实验采用 2017 年 6 月 26 日更新的 OMIM 数据共 4 027 条,表 2 列出了实验中使用数据集的特征.具体地,对应到疾病信息网络中,即涵盖 3 063 个基因节点、4 027 个疾病节点、39 166 个表型节点、3 037 条 *GD*(基因导致疾病)边、4 015 条 *DP*(疾病具有的表型)边.

Table 2 Characteristic of data sets

表 2 数据集特征

基因(个数)	疾病(种)	表型(个数)	基因-疾病(个数)	疾病-表型(个数)
3 063	4 027	39 166	3 037	4 015

gSim-Miner 算法使用 Python 编程实现,Python 版本为 3.6,图数据库采用 Neo4j,版本为 3.2.1,所有实验都在配置为 Intel core i7-3770,3.40 GHz CPU,16GB 内存,Windows10 操作系统的 PC 上完成.gSim-Miner 算法执行参数的默认值为 k=10,即,搜索与给定查询基因具有最大表型相似度的 k 个基因.

同时,为便于存储与准确地计算基因间的表型相似度,我们仅保留了含有 UMLS 和 HPO 注释的表型,表 3 列出了 OMIM 数据中表型术语被不同本体注释的数目.从表 3 的结果也可以看出:UMLS 本体注释涵盖范围最

广,约占所有表型的90.1%.虽然SMEDCT也注释了约8%的表型,但与UMLS本体注释存在大量冗余,如果引入,会降低表型相似度计算的准确性.而ICD9CM和ICD10CM注释的表型约占3%左右,几乎可以忽略.此外,引入HPO的主要目的是修正UMLS通过工具转换时存在的错误,从而使得相关基因更相似、非相关基因差异更大.

Table 3 Number of terms with different ontology annotations

表3 不同本体注释术语的数目

本体库	个数
UMLS	35 292
HPO	4 480
ICD9CM	1 050
ICD10CM	1 451
SNOMEDCT	3 582

此外,图2给出了疾病信息网络构建完成后的可视化结果,其中,用不同颜色节点表示疾病、基因、表型,不同节点间的连线则表示它们之间的关系.图2随机给出了部分疾病、基因与表型的信息,数据规模仅占疾病信息网络的千分之四.由此可见,疾病信息网络的规模十分庞大而且关系极其复杂.

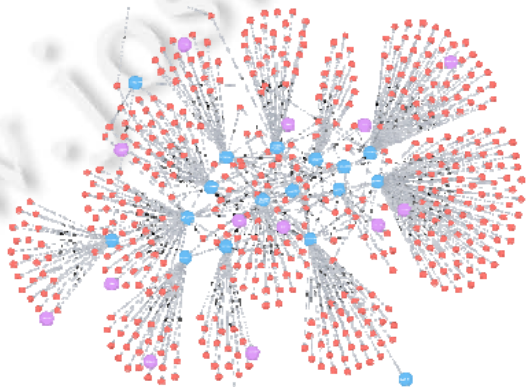


Fig.2 Visualization of disease information network

图2 疾病信息网络可视化

5.2 有效性测试

本文随机从疾病信息网络中抽取了3个代表性基因MC4R,KRAS,GABRB3.表4以“(基因,表型相似度)”的形式列出了与给定查询基因表型最为相似的基因集合.请注意,表型相似度过低的基因不具有临床研究价值,所以我们仅列出基因表型相似度大于0.5的基因集合,即$Sim(g_1,g_2)>0.5$.同时,对于表型相似度相同的情况,我们根据两个基因间符合GDPDG的路径实例数进行降序排序.观察表4可以发现:每一个基因均可以通过计算表型相似度找到自身,即自身相似度最高.此外,以与肥胖症相关的MC4R基因为例,我们挖掘出了PPARG,UCP3,NR0B2,GHRL,ADRB2,ADRB3,UCP1,UCP2,POMC,AGRP,CARTPT等均涉及到肥胖综合征的基因的集合.

Table 4 Genes with top-15 similar phenotype

表4 top-15表型相似基因

基因	top-k个相似基因(k=15)
MC4R	(MC4R,1.0),(PPARG,1.0),(UCP3,1.0),(NR0B2,1.0),(GHRL,1.0),(ADRB2,1.0),(ADRB3,1.0),(UCP1,1.0),(UCP2,1.0),(POMC,1.0),(AGRP,1.0),(CARTPT,1.0),(CEP290,0.64),(NTRK2,0.60),(MOG,0.57)
KRAS	(KRAS,1.0),(TP53,1.0),(MADH4,1.0),(STK11,1.0)
GABRB3	(GABRB3,1.0),(GRIN2B,0.70),(HNRNPU,0.60),(KCNB1,0.57)

为验证gSim-Miner算法设计的有效性,我们对比了仅考虑UMLS或HPO本体注释以及综合考虑二者计算基因表型相似度的搜索结果.具体见表5(同样以“(基因,表型相似度)”形式呈现).由表5可以看出:综合考虑HPO

和 UMLS 不仅可以找到与 MC4R 最相似的基因,同时可以降低一些非相关基因的相似度.原因在于:OMIM 数据中,部分表型仅有 HPO 注释或者仅有 UMLS 注释,单纯考虑某一种本体注释并未将全部表型涵盖,从而减少了基因间仅通过某种本体注释的表型的连通实例,进而降低了基因表型的相似度.而综合考虑可以使得挖掘的基因结果相关的更相似、非相关的差异更大.

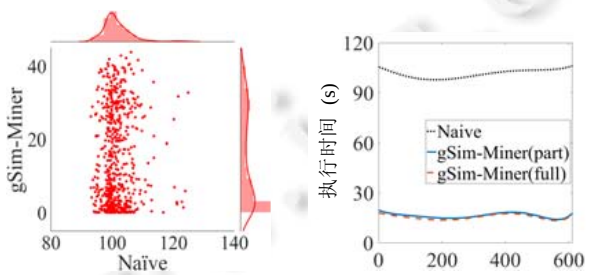
Table 5 Results of considering ontology annotations

表 5 考虑本体注释的结果

MC4R	仅考虑 HPO		仅考虑 UMLS		综合考虑 HPO 和 UMLS	
top-20	(MC4R,1.0)	(PPARG,1.0)	(MC4R,1.0)	(PPARG,1.0)	(MC4R,1.0)	(PPARG,1.0)
	(UCP3,1.0)	(NR0B2,1.0)	(UCP3,1.0)	(NR0B2,1.0)	(UCP3,1.0)	(NR0B2,1.0)
	(GHRL,1.0)	(ADRB2,1.0)	(GHRL,1.0)	(ADRB2,1.0)	(GHRL,1.0)	(ADRB2,1.0)
	(ADRB3,1.0)	(UCP1,1.0)	(ADRB3,1.0)	(UCP1,1.0)	(ADRB3,1.0)	(UCP1,1.0)
	(POMC,1.0)	(UCP2,1.0)	(UCP2,1.0)	(POMC,1.0)	(UCP2,1.0)	(POMC,1.0)
	(AGRP,1.0)	(CARTPT,1.0)	(AGRP,1.0)	(CARTPT,1.0)	(AGRP,1.0)	(CARTPT,1.0)
	(CEP290,0.64)	(IFT74,0.53)	(NTRK2,0.69)	(CEP290,0.64)	(CEP290,0.64)	(NTRK2,0.60)
	(MYT1L,0.53)	(PTHB1,0.52)	(MOG,0.60)	(IFT74,0.59)	(MOG,0.57)	(IFT74,0.57)
	(MOG,0.52)	(BLK, 0.49)	(PTHB1,0.53)	(BBIP1,0.53)	(PTHB1,0.53)	(BLK,0.51)
	(BBS7,0.48)	(NTRK2,0.48)	(BBS12,0.53)	(BLK,0.52)	(BBS7,0.51)	(BBS12,0.49)

5.3 执行效率

本文以疾病信息网络为基础,为了验证执行效率,本文使用两种算法进行对比:Naïve 算法以及包含剪枝策略 1 的 gSim-Miner 算法,记为 gSim-Miner(part);使用剪枝策略 1、策略 2 的 gSim-Miner 算法,记为 gSim-Miner (full).此外,为充分展示剪枝策略对执行时间的影响,我们随机选取了全部基因数据的 20%(612 个基因)用于实验.图 3(a)中,横坐标和纵坐标分别表示 Naïve 算法和 gSim-Miner 算法处理 612 个查询基因的运行时间,同时给出数据的分布情况.观察图 3(a)可知:gSim-Miner 算法执行时间的数据分布较 Naïve 算法分散些,而 Naïve 算法中每个查询基因的执行时间相差无几,大部分集中在[95,105]区间内.原因在于:gSim-Miner 算法因为剪枝策略的影响,降低了候选基因集合个数,所以执行时间产生了比较大的波动.但总体来说,gSim-Miner 算法的执行效率要远高于 Naïve 算法,通过计算,约提高了 6.8 倍.图 3(b)则展示了仅执行剪枝策略 1 以及执行全部剪枝策略的 gSim-Miner 算法与 Naïve 算法的效率对比,图中不同颜色的线条表示不同算法执行 612 个查询基因运行时间的拟合结果.观察可知:剪枝策略 1 可以帮助我们减掉大量候选基因;剪枝策略 2 虽然较剪枝策略 1 对算法执行效率提升不多,但也起到剪枝的作用.



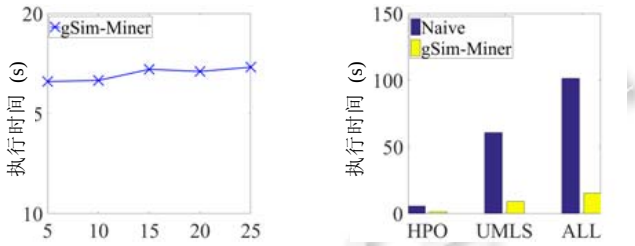
(a) Naïve vs gSim-Miner (b) 剪枝策略对算法影响

Fig.3 Efficiencycomparison between Naïve and gSim-Miner

图 3 对比不同算法的执行效率

图 4(a)、图 4(b)展示了参数 k 和采用不同本体注释对算法执行效率的影响.观察图 4(a)可以发现:随着 k 值的增大,gSim-Miner 算法时间基本呈线性增长,但对整体运行时间变化不大,平均耗时 16.98s.原因在于:gSim-Miner 算法主要开销在于计算基因连通路径的过程以及候选基因的规模,而 k 值对整体执行效率的影响几乎可以忽略.图 4(b)则展示了单独采用 HPO 或者 UMLS 本体注释以及综合考虑二者的情况下对算法的执行效率的

影响,可以发现:Naïve 算法受本体注释的影响较大;而 gSim-Miner 算法执行时间虽随着注释数目的增多也略有提高,但提升幅度很小.结合表 5 数据以及图 4(c)可以看出:我们采用 UMLS 和 HPO 本体注释,不仅可以在一定程度上提高准确度,而且并未花费大量的计算代价.



(a) k 对效率的影响 (b) 本体注释对效率影响

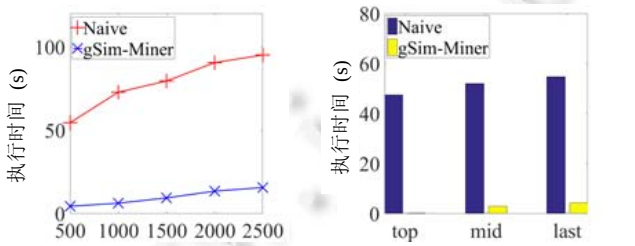
Fig.4 Influence of different parameters on the algorithm

图 4 不同参数对算法的影响

5.4 可扩展性实验

为了验证 gSim-Miner 算法的可扩展性,本文还进行了增加基因节点和扩展基因连通路径数目的实验来比较 Naïve 算法 gSim-Miner 算法与的执行效率.图 5(a)的实验中,我们采用默认的 k 值,变量为基因节点的个数,主要对比在不同节点数目的情况下,Naïve 算法和 gSim-Miner 算法的执行效率.由图 5(a)所示的结果可以看出:当基因节点数目增加时,执行时间会在一定程度上有所提高,但基本呈线性增长,而且增长速度要低于 Naïve 算法,因此也证明了剪枝后的 gSim-Miner 算法可以高效地处理大规模的疾病信息网络.

为了更有效地验证基因连通路径数目改变时对 gSim-Miner 算法可扩展性的影响,本实验通过遍历构建的疾病信息网络,计算出符合 GDP 疾病路径的实例数,即基因通过疾病连通表型的路径实例数并对其进行升序排序,分别取 top300,mid300,last300 对应的基因集合作为测试数据集.从图 5(b)的结果可以明显看出,不同的路径实例数目对算法的执行时间也有一定的影响.分析原因,算法主要的开销之一在于计算不同基因间连通路径实例的数目,而基因到表型的路径实例的数目越多,即基因的出度越多,计算基因间连通路径的规模也就越大,算法整体的执行时间就会越多,但基本维持在线性可控的范围内.



(a) 节点数目对执行效率的影响 (b) 连通路径数目对执行效率的影响

Fig.5 Influence of different conditions on algorithm scalability

图 5 不同条件对算法可扩展性的影响

6 结 论

随着人类对疾病的认识不断加深,如何及早预防和有效地治疗疾病已成为探索的热点.而人类基因组测序工作的完成,也为我们提供了大量与基因相关的生物数据.所以,如何高效地存储和管理大量的生物数据以及如何设计有效的算法和工具去辅助科研工作者挖掘基因疾病数据隐含的未知规律具有重要的意义.然而目前,基于图对 OMIM 数据进行管理的工作还未系统地开展起来,而且关于基因相似性的工作大多是是基于功能关系

和拓扑结构,并未考虑基因与疾病、疾病与表型间的关系.针对这些问题,本文提出了基于疾病信息网络的 top-*k* 相似基因挖掘算法 gSim-Miner,同时设计了高效的剪枝策略.最后,在真实的数据集上验证了 gSim-Miner 算法的有效性、执行效率和可扩展性.

下一步,我们将融合更多的生物医学信息数据库,如解剖数据库、基因工程等,构建出一个更为完善的疾病信息网络,为更多的科研工作者提供分析与计算,从而推动疾病领域研究工作的进展.我们还将根据实际需要,考虑从不同层次对表型进行抽象,从而有助于理解基因在何种类型的表型下拥有共同的作用.此外,我们还将考虑多个协同基因对疾病致病的影响程度.同时,还可尝试将 gSim-Miner 算法与实际的应用场景相结合,以便在实际中验证 gSim-Miner 算法的有效性.

References:

[1] Freimer N, Sabatti C.The human phenome project. Nature Genetics, 2003,34(1):15−21. [doi: 10.1038/ng0503-15]

[2] Oetting WS, Robinson PN, Greenblatt MS, Cotton RG, Beck T, Carey JC, Doelken SC, Girdea M, Groza T, Hamilton CM, Hamosh A, Kerner B, MacArthur JA, Maglott DR, Mons B, Rehm HL, Schofield PN, Searle BA, Smedley D, Smith CL, Bernstein IT, Zankl A, Zhao EY. Getting ready for the human phenome project: The 2012 forum of the human variome project. Human Mutation, 2013, 34(4):661−6. [doi: 10.1002/humu.22293]

[3] Mckusick VA. Mendelian inheritance in man and its online version, OMIM. American Journal of Human Genetics, 2007,80(4): 588−604. [doi: 10.1086/514346]

[4] Sun YZ, Han JW. Mining Heterogeneous Information Networks: Principles and Methodologies. Morgan & Claypool Publishers, 2012. [doi: 10.2200/S00433ED1V01Y201207DMK005]

[5] Sun YZ, Han JW, Zhao PX, Yin ZJ, Cheng H, Wu TY. RankClus: Integrating clustering with ranking for heterogeneous information network analysis. In: Proc. of the 12th Int'l Conf. on Extending Data Base Technology. 2009. 565. [doi: 10.1145/1516360.1516426]

[6] Sun YZ, Yu Y, Han JW. Ranking-Based clustering of heterogeneous information networks with star network schema. In: Proc. of the Int'l Conf. on Knowledge Discovery and Data Mining. 2009. 797−806. [doi: 10.1145/1557019.1557107]

[7] Ji M, Sun YZ, Danilevsky M, Han JW, Gao J. Graph regularized transductive classification on heterogeneous information networks. In: Proc. of the European Conf. on Machine Learning and Principles and Practice of Knowledge Discovery in Database. 2010. [doi: 10.1007/978-3-642-15880-3_42]

[8] Sun YZ, Han JW, Yan XF, Yu PS, Wu TY. PathSim: Meta path-based top-*K* similarity search in heterogeneous information networks. Proc. of the VLDB Endowment, 2011,4(11):992−1003. [doi: 10.2200/S00433ED1V01Y201207DMK005]

[9] Sun YZ, Aggarwal CC, Han JW. Relation strength-aware clustering of heterogeneous information networks with incomplete attributes. Proc. of the VLDB Endowment, 2012,5(5):394−405. [doi: 10.14778/2140436.2140437]

[10] Huang Z, Zheng Y, Cheng R, Sun YZ, Mamoulis N, Li X. Meta structure: Computing relevance in large heterogeneous information networks. In: Proc. of the ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. 2016. 1595−1604. [doi: 10.1145/2939672.2939815]

[11] Digital bibliography & library project. 2017. http://dblp.org/

[12] Chen L, Li X, Han JW. MedRank: Discovering influential medical treatments from literature by information network analysis. In: Proc. of the 24th Australasian Database Conf. Australian Computer Society. 2013. 3−12.

[13] Jeh G, Widom J. Scaling personalized Web search. In: Proc. of the Int'l Conf. on World Wide Web. 2003. 271−279. [doi: 10.1145/775152.775191]

[14] Qi GJ, Aggarwal CC, Huang TS. On clustering heterogeneous social media objects with outlier links. In: Proc. of the Int'l Conf. on Web Search and Web Data Mining. 2012. 553−562. [doi: 10.1145/2124295.2124363]

[15] Rossi RG, Faleiros TDP, Lopes ADA, Rezende SO. Inductive model generation for text categorization using a bipartite heterogeneous network. In: Proc. of the Int'l Conf. on Data Mining. 2012. 1086−1091. [doi: 10.1109/ICDM.2012.130]

[16] Zhang J, Kong X, Jie L, Chang Y, Yu PS. NCR: A scalable network-based approach to co-ranking in question-and-answer sites. In: Proc. of the Int'l Conf. on Information and Knowledge Management. 2014. 709−718. [doi: 10.1145/2661829.2661978]

[17] Ren X, Liu J, Yu X, Khandelwal U, Gu Q, Wang L, Han J. ClusCite: Effective citation recommendation by information network-based clustering. In: Proc. of the Int'l Conf. on Knowledge Discovery and Data Mining. 2014. 821−830. [doi: 10.1145/2623330.2623630]

[18] Alkindy B, Guyeux C, Couchot JF, Salomon M, Bahi JM. Gene similarity-based approaches for determining core-genes of chloroplasts. In: Proc. of the Int'l Conf. on Bioinformatics and Biomedicine. 2015. 71−74. [doi: 10.1109/BIBM.2014.6999130]

[19] Du Z, Li L, Chen CF, Yu PS, Wang JZ. G-SESAME: Web tools for GO-term-based gene similarity analysis and knowledge discovery. Nucleic Acids Research, 2009,37:W345−W349. [doi: 10.1093/nar/gkp463]

[20] Sanfilippo A, Baddeley B, Beagley N, Riensche R, Gopalan B. Enhancing automatic biological pathway generation with GO-based gene similarity. In: Proc. of the Int'l Joint Conf. on Bioinformatics, Systems Biology and Intelligent Computing. 2009. 448−453. [doi: 10.1109/IJCBS.2009.96]

[21] Baralis E, Bruno G, Fiori A. Measuring gene similarity by means of the classification distance. Knowledge & Information Systems, 2011,29(1):81−101. [doi: 10.1007/s10115-010-0374-0]

[22] Othman RM, Deris S, Illias RM. A genetic similarity algorithm for searching the gene ontology terms and annotating anonymous protein sequences. Journal of Biomedical Informatics, 2008,41(1):65−81. [doi: 10.1016/j.jbi.2007.05.010]

[23] Nagar A, Almubaid H. A new path length measure based on GO for gene similarity with evaluation using SGD pathways. In: Proc. of the Int'l Symp. on Computer-Based Medical Systems. 2008. 590−595. [doi: 10.1109/CBMS.2008.27]

[24] Alvarez MA, Yan C. A graph-based semantic similarity measure for the gene ontology. Journal of Bioinformatics & Computational Biology, 2011,9(6):681−695. [doi: 10.1142/S0219720011005641]

[25] Alvarez MA, Qi X, Yan C. A shortest-path graph kernel for estimating gene product semantic similarity. Journal of Biomed Semantics, 2011,2(1):1−9. [doi: 10.1186/2041-1480-2-3]

[26] Webber J. A programmatic introduction to Neo4j. In: Proc.of the 3rd Annual Conf. on Systems, Programming, and Applications: Software for Humanity. 2012. 217−218. [doi: 10.1145/2384716.2384777]

侯泳旭(1994－),女,河北石家庄人,硕士生,CCF 学生会员,主要研究领域为数据挖掘.

卢莉(1972－),女,博士,教授,主要研究领域为数据挖掘,环境医学,GIS.

段磊(1981－),男,博士,副教授,主要研究领域为数据挖掘,生物医学信息分析,进化计算.

唐常杰(1946－),男,教授,博士生导师,CCF 杰出会员,主要研究领域为数据科学.

李岭(1969－),男,博士,教授,博士生导师,主要研究领域为医学遗传学,生物信息学.