·专栏· 文章编号：1000-3428(2015)07-0025-06 文献标识码：A 中图分类号：TP391

基于相似性混合模型的蛋白质交互识别

王宇伟，牛 耘，魏 欧

（南京航空航天大学计算机科学与技术学院，南京 210016）

摘 要： 现有采用机器学习方法的蛋白质交互关系识别系统仅以单句为依据，并且存在标注数据缺乏导致训练集规模小的问题。为此，基于相似性混合模型提出一种新的蛋白质交互识别方法。采用基本的关系相似性（RS）模型做初始判断，利用大规模文本计算单词特征间的相似性，在基本 RS 模型的基础上通过特征聚类方式引入单词相似性模型，从而建立一个混合模型。实验结果表明，该方法能够取得较高且较均衡的精确度和召回率，而单词相似性的引入又进一步提高了 F 值，并且其直接利用已有的交互信息，可避免额外的人工标注。

关键词： 蛋白质交互；关系相似性；单词相似性；K 近邻分类；层次聚类

中文引用格式： 王宇伟，牛 耘，魏 欧. 基于相似性混合模型的蛋白质交互识别[J]. 计算机工程，2015，41(7)：25-30，35.

英文引用格式： Wang Yuwei，Niu Yun，Wei Ou. Identification of Protein-protein Interaction Based on Hybrid Similarity Model[J]. Computer Engineering，2015，41(7)：25-30，35.

Identification of Protein-protein Interaction Based on Hybrid Similarity Model

WANG Yuwei，NIU Yun，WEI Ou

（School of Computer Science and Technology，Nanjing University of Aeronautics and Astronautics，Nanjing 210016，China）

【Abstract】 Current machine learning-based Protein-protein Interaction（PPI）identification systems make predictions solely on evidence within a single sentence and suffer from small training set. In this paper，a hybrid similarity model-based approach is proposed to address these issues. A basic Relational Similarity（RS）model is established to make initial predictions. Word similarity matrices are constructed using a corpus-based approach. A clustering algorithm is applied to group words according to their similarity. The obtained word clusters are introduced to the basic RS model to build a hybrid model. Experimental results show that the basic RS model achieves higher and well-balanced precision and recall，and the introduction of the word similarity model further improves the F-score. This approach makes use of known PPI information，thus releases the burden of manual annotation.

【Key words】 Protein-protein Interaction（PPI）；Relational Similarity（RS）；word similarity；K-nearest Neighbor（KNN）classification；hierarchical clustering

DOI： 10.3969/j.issn.1000-3428.2015.07.005

1 概述

蛋白质是生物细胞最重要的成分，它们通过彼此间的作用完成细胞中的大部分过程，蛋白质交互（Protein-protein Interaction，PPI）是生物学研究的重要内容，也是解决大量医学难题的关键信息。因而，生物医学领域专家手工地从医学文献中收集这些信息录入统一格式的数据库中，如 HPRD[1]，IntAct[2]，MINT[3]等。然而随着生物医学文献的急剧增长，手工的方式显然远不能满足实际需要。为帮助生物领域专家有效地从这些文本中获取相应的医学信息，基于自然语言处理的蛋白质交互自动识别技术已成为一项重要的研究内容。本文提出一种基于相似性混合模型的蛋白质交互识别方法，将现有的 PPI 数据库作为训练数据，以避免额外的人工标注负担，并且充分利用大规模文本库资源，依据文本中丰富的上下文信息，更全面地获取交互关系特征。

2 相关工作

目前，从医学文本中自动识别蛋白质交互信息的技术主要包括：基于同现的方法[4]，基于模式匹配、规则的方法[5]和基于机器学习的方法[6-7]。基

基金项目： 国家自然科学基金资助项目（61202132，61170043）。

作者简介： 王宇伟（1989－），男，硕士研究生，主研方向：自然语言处理；牛 耘、魏 欧，副教授、博士。

收稿日期： 2014-08-05 **修回日期：** 2014-08-29 **E-mail：** yniu@nuaa.edu.cn

于同现的方法通过统计 2 个蛋白质在句子中的共现次数来判断蛋白质间是否存在交互关系,该方法简单,但结果召回率较高但精确度低[8];基于模板匹配的方法首先建立能够刻画蛋白质交互关系的句法模式,然后通过文本匹配寻找对应交互关系,该方法理论简单,但难以处理复杂的句子,其有限的覆盖面导致了比较低的召回率,另外手动建立模式需要巨大开销且通常只适用于特定的部分数据[9]。因而一些基于规则的 PPI 抽取方法涌现出来,例如文献[10]提出的规则基于句子句法结构中的依赖关系。文献[11]利用带有语法产生规则的句法分析器来识别 PPI。这些系统着眼于分析整个句子的语法特点,从而充分揭示句中成分之间的关系,能获得更高的准确性,但需要更高的计算能力和时间复杂度。

近年来,越来越多的 PPI 识别技术采用了基于机器学习的方法,这些方法主要有 2 类:基于特征的方法[12]和基于核的方法[13-14]。基于特征的方法从标注有交互关系的句子中抽取重要特征,包括词汇特征,语法特征和语义特征建立模型来判断蛋白质之间的交互关系[15]。基于核的方法通过设计核函数进一步利用句子结构表示(如字符串序列、句法依赖或句法分析)上的隐含特征[16]。然而,目前基于机器学习的方法主要以单句为依据判别一个句子中出现的蛋白质对之间是否存在交互关系。这种方式的优势在于能够在句子级别上提供蛋白质交互的描述和证据,然而它也存在着局限性。首先,在进行交互关系判定时仅依赖于一句话的信息。由于蛋白质交互信息描述语言的多样性和句法结构的复杂性,以单句为依据的方法难以对交互特征进行全面的把握。其次,对蛋白质关系描述的一句话中往往包含 3 对以上蛋白质。而以单句为依据的方法在建立训练集时要求对一个句子中出现的每一对蛋白质都标注其是否交互,这就使得标注训练集代价高昂。正是由于此,采用这种方法的分类模型一般都建立在很小的训练集上,而这必然影响到模型的推广应用的效果。

针对以上这些问题,本文提出了一种基于相似性混合模型的蛋白质交互识别方法。在之前的工作中[17],笔者通过对蛋白质签名档的分析提取一元词特征建立关系相似性(Relational Similarity,RS)模型,取得了比支持向量机模型(Support Vector Machine Model,SVM)更高的 F 值。然而短语结构特征及依赖关系特征的引入并未提高 PPI 的识别精度。与此不同,本文从降低向量的稀疏性从而增加有交互关系蛋白质对的相似性角度出发,利用大规模语料计算出单词相似性以反映出单词特征语义上的联系,并将其引入基本关系相似性模型建立一个新的混合模型。

3 基于 RS 基本模型的 PPI 识别

基于 RS 基本模型的 PPI 识别步骤如下:(1)在医学文献数据库中搜索包含目标蛋白质对的句子集合作为此目标蛋白质对的签名档;(2)从签名档中提取特征,并且把每个蛋白质对实例映射为一个 n 维的特征向量,建立向量空间模型;(3)通过计算向量之间的相似性来判断目标蛋白质对之间是否存在交互关系。

3.1 关系获取

PubMed 数据库收录了超过 1 800 000 篇生物医学文献摘要,是建立 PPI 网络重要的信息来源。对于每一个目标蛋白质对(*Protein*1,*Protein*2),本文在 PubMed 数据库中通过搜索同时包含 *Protein*1 和 *Protein*2 这两个蛋白质的句子集合作为目标蛋白质对的签名档。由于 PubMed 没有提供直接检索句子的接口,本文将其分为以下两步来完成:

(1)在 PubMed 数据库中检索出同时包含目标蛋白质对 *Protein*1 和 *Protein*2 的摘要;

(2)在第(1)步搜索出的摘要中找出同时包含 *Protein*1 和 *Protein*2 的句子。

最后,每一个蛋白质对都会有一个句子集合与之对应,也就是它的签名档。建好蛋白质对的签名档之后,即可以利用这些上下文信息对目标蛋白质对的交互关系做出判断。

3.2 关系表示

本文采用向量空间模型来表示蛋白质 *Protein*1 和 *Protein*2 之间的关系 R。向量的维是刻画这一关系的单词特征,因目标蛋白质对的签名档中包含了关系 R 较完整的描述,把所有签名档中的单词去除那些停止词,单字符单词以及无意义的数字之后剩余的单词作为特征,并且排除了那些低频词(出现此单词的签名档数少于 25 个),最终留有 4 867 个单词特征。每一个关系 R 用一个 4 867 维的特征向量来刻画,这个向量的特征权重采用 2 种方式表示:二值权重(0/1)对应于特征单词是否在目标蛋白质对对应的签名档中出现,出现时特征值为 1,否则就为 0;TF-IDF 权重 w_i 采用式(1)计算:

$$w_i = tf_i \times \lg\left(\frac{N}{df_i}\right) \tag{1}$$

其中,w_i 是第 i 个特征的特征权值;tf_i 是第 i 个特征在签名档中出现的次数;df_i 是指签名档集合中出现第 i 个特征的签名档数目;N 是总的签名档数目。

3.3 关系相似性计算

关系相似性计算通过比较目标蛋白质对与已知

的蛋白质对(包括有交互关系和无交互关系两类)的相似性来判定目标蛋白质对是否具有交互关系。目标蛋白质对以及已知交互关系的蛋白质对分别用 2 个向量 $\boldsymbol{v}_1$,$\boldsymbol{v}_2$ 表示,它们之间的相似性采用余弦值来度量,如下所示:

$$\boldsymbol{v}_1 = (v_{1,1}, v_{1,2}, \cdots, v_{1,i}, \cdots, v_{1,n})$$

$$\boldsymbol{v}_2 = (v_{2,1}, v_{2,2}, \cdots, v_{2,i}, \cdots, v_{2,n})$$

$$\mathrm{sim}(\boldsymbol{v}_1, \boldsymbol{v}_2) = \frac{\sum_{k=1}^{n} v_{1,k} \cdot v_{2,k}}{\sqrt{\sum_{k=1}^{n} (v_{1,k})^2} \cdot \sqrt{\sum_{k=1}^{n} (v_{2,k})^2}} \quad (2)$$

3.4　K 近邻分类

得到了实例的相似性之后,基于相似性采用 K 近邻分类(K-nearest Neighbor,KNN)方法识别目标蛋白质之间的交互关系。KNN 是基于统计的分类方法,也是文本分类中比较常用的方法[18-19],首先查询训练数据中与目标蛋白质对(关系相似性)最相似的 K 个蛋白质对实例,这 K 个最相似的实例中哪种类别的实例最多,就将目标蛋白质对分为哪一类。在此算法中,如果存在多个距离目标蛋白质对一样近的实例,则只保留一个,且这个实例类别取这多个实例中占多数的类别。

3.5　实验数据及设置

实验全部的训练数据来自于现有的 PPI 数据库而不需要额外的人工标注。本文把有交互的蛋白质对看作正样例,无交互的看作负样例。正样例来源于专家手工收集的人类 PPI 数据库 HPRD[1],抽出其中那些包含在 PubMed 数据库 1 篇以上摘要中的蛋白质对作为有交互的蛋白质对训练集,共 1 420 对。而对于负样例,采用生物信息学领域常用的方法,首先对 HPRD 中的蛋白质随机组合成蛋白质对,并且去除包含在 HPRD 数据库中的组合,最后只保留那些包含在 PubMed 数据库一篇以上摘要中的组合作为最后无交互的蛋白质对训练集,共 1 353 对,因此,实验数据集中共包含了 2 773 个蛋白质对。

实验采用的结果性能评价指标是当前 PPI 抽取系统主要使用的 3 个指标:精确率 $P = TP/(TP + FP)$,召回率 $R = TP/(TP + FN)$,F 值 $(F = 2P \times R/(P + R))$。对于数据集中的每个蛋白质对都基于它们的签名档建立对应的特征向量,并采用留一交叉验证法(leave-one-out)进行测试,即将每个蛋白质对作为测试样例,其余的作为训练样例,这样共测试 2 773 次。最后采用 KNN 分类算法完成识别过程。

3.6　识别结果及分析

表 1 列出了以 0/1 为权值的基本关系相似性模型在 K 取不同值时的蛋白质交互识别结果。

表 1　以 0/1 为权重的基本 RS 模型识别结果　%

K	有交互的蛋白质对			无交互的蛋白质对		
	精确率	召回率	F 值	精确率	召回率	F 值
1	75.6	74.5	75.0	73.6	74.7	74.2
3	77.6	73.9	75.7	73.9	77.7	75.7
5	77.8	73.0	75.3	73.4	78.1	75.7
7	77.6	74.1	75.8	74.0	77.6	75.8

从以上结果可以看出,以大规模文本为依据的关系相似性模型采用 0/1 作为权值时取得了较高的 F 值且随着 K 的变化也稳定在 75.5% 左右,识别结果的准确度和召回率也比较均衡,表明描述蛋白质交互关系的文本间存在着共性,即描述交互关系的不同句子对目标蛋白质对的表达是相似的,而这种相似性能够被本文基于相似性的模型有效地捕获从而做出正确的判断。表 2 列出了蛋白质对特征向量以 TF-IDF 值为权重采用 KNN 分类方法的识别结果,由结果可以看出,以 TF-IDF 为权重的识别结果的精确度和召回率相差较大且 F 值也低于表 1 中以 0/1 为权值的基本关系相似性模型的识别结果。在接下来混合模型的实验中均以0/1作为权值。

表 2　以 TF-IDF 为权重的基本 RS 模型识别结果　%

K	有交互的蛋白质对			无交互的蛋白质对		
	精确率	召回率	F 值	精确率	召回率	F 值
1	65.8	81.3	72.7	73.9	55.6	63.5
3	65.4	83.0	73.2	75.2	53.9	62.8
5	64.3	83.8	72.8	75.1	51.1	60.8
7	64.2	85.1	73.2	76.2	50.3	60.6

4　基于混合模型的 PPI 识别

在上文所提出的基本关系相似性模型中,存在单词特征向量的高维、稀疏等问题,即只有少量特征在特定蛋白质对的签名档中出现。另外基本关系相似性模型忽略了特征单词之间的语义相关性使得相似性计算结果偏低。如 bind 和 interact 是向量空间模型中的特征,而这两个单词可能分别只单独出现在交互蛋白质对 A 和 B 的签名档中,A 的签名档中只出现了 bind,而 B 的签名档中只出现了 interact,因此,对应 A 和 B 的特征向量中 bind 和 interact 的特征值只有一个非 0。然而这两个特征都描述了蛋白质之间的交互关系,两者之间存在着语义相关性,这种相关性在基本模型中被忽略从而会导致 A 与 B 的关系相似度值偏低。针对这种现象,首先基于大规模语料计算出单词特征间的相似性,而后采用聚类的方式将相似性结果引入基本关系相似性模型中构建了新的混合模型从而改进关系相似性的计算,进一步提高蛋白质交互的识别精度。

4.1 单词相似性模型

4.1.1 基于语料的单词相似性计算

本文采用基于语料的方法来计算单词间的相似性,即认为相似的 2 个单词会出现在相似的上下文环境当中[20],因此,通过比较上下文的相似性可以得到单词的相似性。把目标单词用其在大规模语料中的上下文特征来表示,上下文特征为目标单词周围一定大小窗口内的所有单词。在本文实验中,PubMed 中的医学文本摘要被作为大规模语料的来源。

4.1.2 目标词集合

本文把所有蛋白质对的签名档中的单词去除停止词,单字符单词及无意义的数字作为初始目标词,这些词也是第 3 节基本模型中蛋白质对特征的来源(3.2 节)。根据其在签名档中词性标注信息对其分组,分为名词组(noun)、动词组(verb)、形容词组(adjective)和副词组(adverb),一个单词可能出现在多个词性组中。只有这 4 组中的单词被作为最终的目标词进行词相似性计算,其余词性的单词被去除。为得到单词词性(Part-of-speech, POS),本文采用 Apache OpenNLP 句法分析工具(http://opennlp. apache. org/index. html)对签名档中的句子做了词性标注。下面计算每个分组内两两目标词间的相似度。

4.1.3 单词相似性矩阵

本文对每组内两两单词间的相似性进行计算得到对应的 4 个单词相似性矩阵,以动词组 *GV* 为例,其计算过程如下:

步骤 1 获取 *GV* 中目标词的上下文

从 PubMed 数据库随机抽取 1 GB 大小的文本摘要作为语料,语料中与一个目标词相邻的左、右各 5 个单词去除停用词,单字符单词及无意义的数字后剩余的单词作为这个目标词的上下文。

步骤 2 为目标词建立共现特征矩阵

GV 中所有目标词的上下文单词被作为目标词的共现向量特征,特征权值采用条件概率 $P(w|w_1)$,其中,w_1 是目标词;w 是特征词。其值为语料中 w 和 w_1 在设定大小窗口内共现的次数除以 w_1 在整个语料库中出现的次数。最后建立目标词的共现矩阵 A,每一行是 *GV* 中一个目标词的共现特征向量。

步骤 3 计算同组中两两单词间的相似度

单词间相似性的度量采用 2 个单词特征向量的余弦距离(式(2)),*GV* 中两两单词间相似性的计算结果形成相似性矩阵 B,其第 i 行第 j 列元素 b_{ij} 是对应的第 i 个目标单词与第 j 个目标单词间的相似度。将共现矩阵 A 行单位后,用 A 乘以 A 的转置 A^{T} 即得到矩阵 B。

表 3 是单词相似性计算的结果示例,第 1 列为目标词,第 2 列是与第 1 列目标词最相似的其他单词及对应的相似度,按照相似度降序排列。

表 3 单词相似性计算结果示例

目标词	最相似的单词及对应的相似度
binds	(interacts, 0. 846 49), (binding, 0. 744 41), (bind, 0. 744 36), (bound, 0. 710 35), (interacte, 0. 705 52), (interact, 0. 700 62)
enhances	(enhanced, 0. 907 21), (enhance, 0. 85 55), (augment, 0. 825 22), (inhibits, 0. 820 10), (augmented, 0. 817 56), (suppresses, 0. 802 75), (affect, 0. 802 26)
examine	(investigate, 0. 961 31), (determine, 0. 911 81), (evaluate, 0. 901 32), (explore, 0. 897 00), (assess, 0. 859 86), (ascertain, 0. 840 71)

从计算结果可以看出,基于语料的方法能有效地计算出单词间的相似性,并能够有效地反映出2 个单词对于特定领域内的语义上的联系,如:bind, interact 和 activate 在蛋白质交互领域有着语义上很大的相似性,它们都是用来描述 PPI 的典型单词。

4.2 基于混合模型的 PPI 识别

本节采用特征聚类的方式将 4.1 节计算出的 4 个单词相似性矩阵引入到基本关系相似性模型中建立混合模型。在这个混合模型中,先用以 0/1 为权值的基本的关系相似性模型进行 1NN 分类及 $K(K>1)$NN分类,如果利用两者分类的结果一致,其分类结果将作为最终的识别结果,否则根据单词相似性对单词进行聚类,利用聚类的结果重新生成蛋白质对的特征向量,最后再把新特征向量的 1NN 分类结果作为混合模型的识别结果。

4.2.1 4 组目标词的聚类

对于每组目标词,本文根据 4.1 节得到的相似性矩阵做层次聚类。层次聚类算法又称为树聚类算法[21-22],通过对单词集按照分裂或聚合的方式进行层次分解,以形成一个层次序列。以聚合方式进行层次聚类称为凝聚层次聚类,具体描述下:

算法 1 凝聚层次聚类

输入 a 个单词及单词相似性矩阵 $\boldsymbol{M}_{a\times a}$

输出 层次聚类树

初始:将每个单词作为一个聚类簇

Repeat:

(1)根据相似性矩阵合并最接近的 2 个簇;

(2)更新相似性矩阵,以反映新合并得到的这个簇与原来簇之间的相似性。

Until 仅剩下一个簇

本文使用 Matlab 的层次聚类工具(http://www.mathworks.cn/cn/help/stats/hierarchical-clustering-2.html)完成该算法的聚类过程,2 个簇之间的相似性使用的是组平均技术,它定义簇的相似度取自 2 个簇中所有点对相似度的平均值。聚类簇数目的结果如表 4 所示。4 个词性的单词组中动词(verb)有 1 857 个,名词

(noun)3 564 个,形容词(adj)2 253 个,副词(adv)361 个。表 4 中参数 Z 是切割层次聚类树得到聚类簇的一个阈值,其取值决定了聚类簇数目的多少,如把 1.0 作为阈值,动词被聚成了 643 个类簇,以 0.9 作为阈值,动词被聚成 709 个类簇。

表 4　聚类数目结果

Z	verb	noun	adj	adv
0.7	1 264	2 418	1 566	254
0.8	750	1 450	974	151
0.9	709	1 352	915	143
1.0	643	1 212	830	136

4.2.2　混合模型的识别算法

将单词凝聚层次聚类用于 PPI 的识别过程,具体描述如下:

算法 2　混合模型 PPI 识别

输入　蛋白质对训练集及它们的签名档集;蛋白质对单词特征集 $F=\{f_n\}_{1\leqslant n\leqslant N}$;待识别的目标蛋白质对 P 及其签名档

输出　识别蛋白质对 P 的交互关系结果:有交互或无交互

(1)特征集变换(具体过程如图 1 所示)

1)将特征集 F 中的单词根据其在签名档中的词性标注信息分到对应的词性组中,若其不属于任何一个词性组,则将其单独作为一组。

2)将词性组内的单词采用层次聚类法聚成类簇,只有一个单词的分组作为一个聚类簇,每个簇作为一个特征,最终所有的聚类簇形成新的特征集 F'。

(2)蛋白质对交互关系识别

1)将待识别的蛋白质对 P 及训练集的每个蛋白质对用特征集 F 上的权值为 0/1 的特征向量表示(见 3.2 节)。

2)对 P 的特征向量采用 1NN 及 $KNN(K>1)$ 分别作初始分类。

3)如果两者分类结果一致,则将其作为 P 的混合模型识别结果。

4)否则:

①对于训练集中每个蛋白质对及目标蛋白质对 P 的签名档,在新的特征集 F' 上生成特征向量表示,表示方法如下:若特征簇中有一个单词在蛋白质对的签名档中出现,则对应簇特征的特征权值为 1,若特征簇中所有单词均未在蛋白质对的签名档中出现,则该簇特征的特征权值为 0。

②对 P 的新特征向量,用 1NN 分类结果作为其混合模型的识别结果。

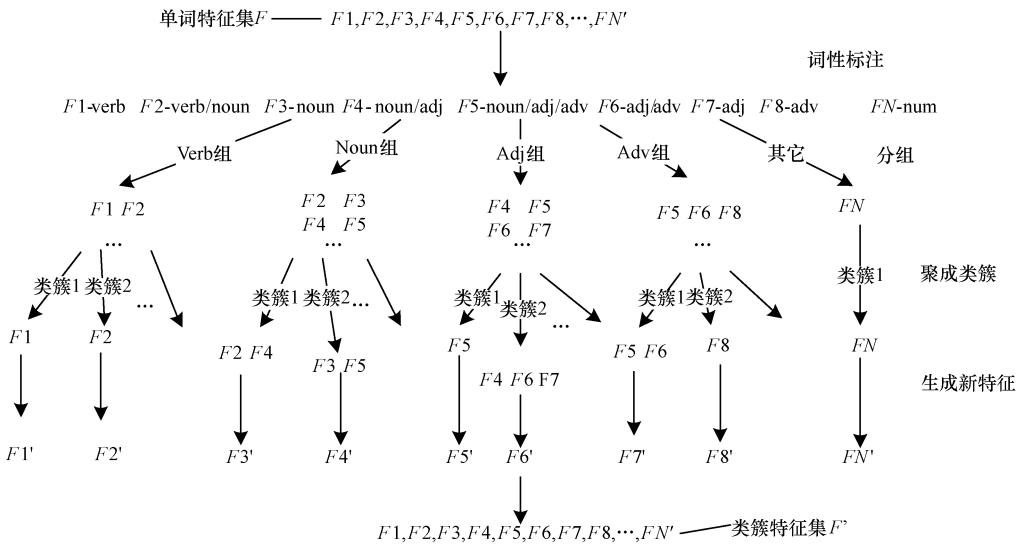

图 1　以类簇为特征的新特征集生成过程

4.3　识别结果及分析

本节比较了混合模型与基本关系相似性模型的识别结果,表 5、表 6 分别显示了混合模型识别有交互蛋白质对和无交互的蛋白质对在 K 和聚类簇数目取不同值时的结果,其中对于每一个 K 值组,3 个评估结果各自的最大值都以黑体显示。为进行对比,本文同时给出基本 RS 模型的识别结果。可以看出,混合模型的全部结果均高于基本模型。当 K 为 7 时,混合模型取得的 F 值最高,识别正例蛋白质对的 F 值比基本模型高 2.5% ,负例高出约 3 个百分点。这清晰地表明单词相似性的引入产生了更优的模型。对于正例蛋白质对,精确度当 Z 取 1 时,得到最好值,而召回率和 F 值则在 Z 取 0.8 时取得最佳。对于负例蛋白质对,与此相反,对于不同的 K,精确度都是在 $Z=0.8$ 时最佳,而召回率和 F 值在 $Z=1$ 时最好。

表 5 有交互蛋白质对的识别结果 %

K	Z	精确率	召回率	F 值
3	0.7	76.9	**75.6**	76.2
	0.8	78.0	**75.6**	**76.8**
	0.9	77.4	**75.6**	76.5
	1.0	**78.4**	75.0	76.6
5	0.7	77.5	75.6	76.5
	0.8	**78.3**	**76.1**	**77.1**
	0.9	77.8	75.8	76.8
	1.0	**78.7**	75.4	77.0
7	0.7	77.1	76.1	76.6
	0.8	78.3	**76.7**	**77.5**
	0.9	78.0	76.1	77.0
	1.0	**78.8**	76.0	77.4
基本 RS 模型		75.6	74.5	75.0

表 6 无交互蛋白质对的识别结果 %

K	Z	精确率	召回率	F 值
3	0.7	74.8	76.2	75.5
	0.8	**75.2**	77.6	76.4
	0.9	75.0	76.9	75.9
	1.0	74.9	**78.3**	**76.5**
5	0.7	75.0	76.9	76.0
	0.8	**75.6**	77.8	76.7
	0.9	75.3	77.3	76.3
	1.0	75.3	**78.6**	**76.9**
7	0.7	75.2	76.3	75.7
	0.8	**76.1**	77.8	76.9
	0.9	75.6	77.5	76.5
	1.0	75.7	**78.6**	**77.1**
基本 RS 模型		73.6	74.7	74.2

5 结束语

本文基于大规模文本建立了蛋白质交互识别的关系相似性模型,针对基本关系相似性模型特征稀疏及忽略特征单词间相关性的问题,采用基于大规模语料的方法计算出单词特征间的相似性矩阵,以单词聚类方式将单词相似性模型引入基本模型中。从实验识别结果可以看出,基于大规模文本的方法自动识别 PPI 能够取得较高且较均衡的精确度和召回率,并且单词相似性的引入也进一步提高了模型的识别精度。与广泛采用的基于单句的机器学习方法不同,本文方法直接以蛋白质对为研究对象,以大规模文本作为识别依据,识别的结果可直接用于 PPI 网络的构建,并且还能充分利用已有的 PPI 数据而无需额外的人工标注。

参考文献

[1] Prasad T S K,Goel R,Kandasamy K,et al. Human Protein Reference Database——2009 Update[J]. Nucleic Acids Research,2009,37(s1):767-772.

[2] Kerrien S,Alam-Faruque Y,Aranda B,et al. IntAct——Open Source Resource for Molecular Interaction Data[J]. Nucleic Acids Research,2007,35(s1):561-565.

[3] Ceol A,Aryamontri A C,Licata L,et al. MINT,the Molecular Interaction Database:2009 Update[J]. Nucleic Acids Research,2010,38(s1):532-539.

[4] Bunescu R,Mooney R,Ramani A,et al. Integrating Co-occurrence Statistics with Information Extraction for Robust Retrieval of Protein Interactions from Medline[C]// Proceedings of Workshop on LinkingNatural Language Processing and Biology:Towards Deeper Biological Literature Analysis.[S.l.]:Association for Computational Linguistics,2006:49-56.

[5] Koike A,Kobayashi Y,Takagi T. Kinase Pathway Database:An Integrated Protein-kinase and NLP-based Protein-interaction Resource[J]. Genome Research,2003,13(6A):1231-1243.

[6] 杨志豪,洪 莉,林鸿飞,等. 基于支持向量机的生物医学文献蛋白质关系抽取[J]. 智能系统学报,2008,3(4):361-369.

[7] 崔宝今,林鸿飞,张 霄. 基于半监督学习的蛋白质关系抽取研究[J]. 山东大学学报:工学版,2009,39(3):16-21.

[8] Grimes G R,Wen T Q,Mewissen M,et al. PDQ Wizard:Automated Prioritization and Characterization of Gene and Protein Lists Using Biomedical Literature[J]. Bioinformatics,2006,22(16):2055-2057.

[9] Ananiadou S,Kell D B,Tsujii J. Text Mining and Its Potential Applications in Systems Biology[J]. Trends in Biotechnology,2006,24(12):571-579.

[10] Fundel K,Küffner R,Zimmer R. RelEx_Relation Extraction Using Dependencyparse Trees[J]. Bioinformatics,2007,23(3):365-371.

[11] Temkin J M,Gllder M R. Extraction of Protein Interaction Information from Unstructured Text Using a Context-tree Grammar[J]. Bioinformatics,2003,19(16):2046-2053.

[12] Qian Weizhong,Fu Chong,Cheng Hongrong. Semi-supervised Method for Extraction of Protein-protein Interactions Using Hybrid Model[C]//Proceedings of the 3rd International Conference on Intelligent System Design and Engineering Applications. Washington D. C.,USA:IEEE Computer Society,2013:1268-1271.

[13] Zhang Shaowu,Hao Liyang,Zhang Tinghe. Prediction of Protein-protein Interaction with Pairwise Kernel Support Vector Machine[J]. International Journal of Molecular Sciences,2014,15(2):3220-3233.

[14] Chen Peng,Guo Jianyi,Yu Zhengtao,et al. Protein-protein Interaction Extraction Based on Convex Combination Kernel Function[J]. Journal of Computer and Communications,2013,1(5):9-13.

[15] Otasek D,Jurisica I,Niu Yun. Evaluation of Linguistic Features Useful in Extraction of Interactions from PubMed; Application to Annotating Known,High-throughput and Predicted Interactions in I2D[J]. Bioinformatics,2010,26(1):111-119.

(下转第 35 页)

表 4　特征删减后的预测性能对比

删除的特征	准确率	召回率	F1 值
粉丝数	0.768	0.770	0.769
关注数	0.700	0.703	0.700
微博数	0.768	0.770	0.769
认证类型	0.730	0.729	0.729
用户等级	0.768	0.770	0.769
用户发博频率	0.675	0.680	0.677
微博转发速率	0.658	0.667	0.661

当保留全部特征时,模型进行预测的准确率、召回率和 F1 值分别为 0.768,0.770,0.769,进行对比后可以直观地看出不同特征的重要程度:用户粉丝数和微博数特征的删除并未导致性能下降,与本文预期结果相符。根据本文热度指数的定义,拥有大量粉丝和微博数的明星类用户与其他用户在预测时应公平对待,因此,粉丝数和微博数对最终结果无影响也验证了热度指数定义的合理性。而加入用户的发博频率和用户前期的转发速率这 2 项动态特征对模型性能的提升达到约 10%,说明了用户的活跃度和微博初期的受欢迎度对于预测微博热度是非常重要的指标。

5　结束语

本文以新浪微博作为研究对象,通过自主开发的 BigData 爬虫开放平台获取了大量用户、微博及转发数据。利用统计微博样本的转发规律对微博热度进行定义,该热度定义可以作为预测新浪微博热门话题排行的新算法。在现有研究的基础上加入微博动态特征用于预测转发数在 100 以上的微博达到特定热度级别的概率。由于所选用的特征只局限于数量特征,而未考虑微博本身的话题特征,因此准确率还有待提高。下一步将提取微博的文本特征(如主客观态度、微博主题词、命名实体、情感极性等)作为待选特征进行热度预测。

参考文献

[1] Bandari R, Asur S, Huberman B A. The Pulse of News in Social Media: Forecasting Popularity [C]// Proceedings of the 6th International AAAI Conference on Weblogs and Social Media. Palo Alto, USA: AAAI Press, 2012: 26-33.

[2] 杨于峰,余伟萍,田　盼. 基于 SOM 神经网络的品牌丑闻微博传播分类预测研究[J]. 情报杂志, 2013, 32(10): 23-28.

[3] Weng J, Lim E, Jiang J, et al. TwitterRank: Finding Topic-sensitive Influential Twitterers [C]//Proceedings of International Conference on Web Search and Data Mining. New York, USA: ACM Press, 2010: 261-270.

[4] Naveed N, Gottron T, Kunegis J, et al. Bad News Travel Fast: A Content-based Analysis of Interestingness on Twitter[C]//Proceedings of the 3rd International Web Science Conference. New York, USA: ACM Press, 2011: 45-53.

[5] Suh B, Hong L, Pirolli P, et al. Want to be Retweeted Large Scale Analytics on Factors Impacting Retweet in Twitter Network [C]//Proceedings of the 2nd International Conference on Social Computing. Washington D. C., USA: IEEE Press, 2010: 177-184.

[6] Yang J, Counts S. Predicting the Speed, Scale, and Range of Information Diffusion in Twitter [C]// Proceedings of the 4th International AAAI Conference on Weblogs and Social Media. Palo Alto, USA: AAAI Press, 2010: 355-358.

[7] Szabo G, Huberman B. Predicting the Popularity of Online Content[J]. Communications of the ACM, 2010, 53(8): 80-88.

[8] Petrovic S, Osborne M, Lavrenko V. RT to Win! Predicting Message Propagation in Twitter [C]// Proceedings of the 5th International AAAI Conference on Weblogs and Social Media. Palo Alto, USA: AAAI Press, 2011: 586-589.

[9] Hong Liangjie, Dan O, Davison B D. Predicting Popular Messages in Twitter [C]//Proceedings of the 20th International Conference Companion on World Wide Web. New York, USA: ACM Press, 2011: 57-58.

[10] 张　旸,路　荣,杨　青. 微博客中转发行为的预测研究[J]. 中文信息学报, 2012, 26(4): 109-114.

[11] 李英乐,于洪涛,刘力雄. 基于 SVM 的微博转发规模预测方法[J]. 计算机应用研究, 2013, 30(9): 2594-2597.

[12] 黄英来,孙晓芳,刘镇波,等. 微博转发预测算法评测系统的建立及性能比较[J]. 哈尔滨理工大学学报, 2013, 18(4): 52-57.

[13] 熊小兵,周　刚,黄永忠,等. 新浪微博话题流行度预测技术研究[J]. 信息工程大学学报, 2012, 13(4): 496-502.

[14] 刘晓娟,尤　斌,张爱芸. 基于微博数据的应用研究综述[J]. 情报杂志, 2013, 32(9): 39-45.

编辑　陆燕菲

(上接第 30 页)

[16] 唐　楠,杨志豪,林鸿飞,等. 基于多核学习的医学文献蛋白质关系抽取[J]. 计算机工程, 2011, 37(10): 184-186.

[17] 封二英,牛　耘,魏　欧,等. 基于关系相似性的蛋白质交互自动识别[J]. 计算机科学, 2013, 40(6): 229-232.

[18] 许　幸,张启蕊. 基于 KNN 算法的医药信息文本分类系统的研究[J]. 计算机技术与发展, 2009, 19(4): 206-209.

[19] 王　煜,白　石,王正欧. 用于 Web 文本分类的快速 KNN 算法[J]. 情报学报, 2007, 26(1): 60-64.

[20] Mohammad S, Hirst G. Distributional Measures of Semantic Distance: A Survey[Z]. 2012.

[21] Marques J P. 模式识别——原理、方法与应用[M]. 2 版. 吴逸飞,译. 北京:清华大学出版社, 2002.

[22] Fred A L N, Leitao J M N. Partitional vs Hierarchical Clustering Using a Minimum Grammar Complexity Approach[C]//Proceedings of SSPR&SPR'00. Berlin, Germany: Springer, 2000: 193-202.

编辑　金胡考