第42卷 第2期 西 安 交 通 大 学 学 报 Vol.42 №2
2008年2月 JOURNAL OF XI'AN JIAOTONG UNIVERSITY Feb. 2008

结合受控词汇表的生物基因本体标注与分类

崔舒宁，朱丹军，冯博琴，昂正全
（西安交通大学电子与信息工程学院，710049，西安）

摘要：通过研究有关基因的生物学文献特征，提出了一种能对生物基因文献进行自动标注与分类的方法.在 K 最邻近算法的基础上，采用了 Chi-Square 特征选择方案，并且在加权算法中突出了 Chi-Square 的选择特点.另外，采用文档逻辑分块法，将额外的生物受控词汇表中的信息所形成的向量直接引入到了分类算法中，以提高分类和标注的效果.实验表明，所提算法优于常用的单词频率/逆文档频率加权方法，其在文本检索大会（TREC）数据集上的分类、标注效果分别比 TREC 公布的最好结果提高了 3.14%和 4.12%.

关键词：基因本体；分类标注；最邻近算法
中图分类号：TP319 **文献标志码**：A **文章编号**：0253-987X(2008)02-0171-04

Triage and Annotation of Biological Gene's Ontology Combined Controlled Glossary

CUI Shuning, ZHU Danjun, FENG Boqin, ANG Zhengquan
(School of Electronics and Information Engineering, Xi'an Jiaotong University, Xi'an 710049, China)

Abstract: Based on the K nearest neighbor algorithm, an improved method was proposed for selecting genes-related documents from biology literature, and then automatically annotating and classifying. The method employs the Chi-Square feature selection plan and highlights the Chi-Square selections in weighted calculations. Furthermore, the effect of classification and annotation was improved by dividing the documents into logical blocks and introducing additional vectors from biological resources MeSH into the classification algorithm directly. Experiment results show that the proposed method is better than the commonly used TFIDF (term frequency and inverse document frequency) weighting method, and the results tested on TREC (text retrieval conference) data sets are 3.14% higher in classification and 4.13% higher in annotation comparing to the best results announced TREC.

Keywords: gene ontology; classification annotation; nearest neighbor algorithm

在生物学文献中，同一物质以不同的书写方式出现在文档集合中的现象比较明显.生物术语学研究者建立了一系列受控词汇表或本体来解决此类问题，如美国国家医学图书馆开发的受控词汇表 MeSH(medical subject headings)、统一医学语言系统(unified medical language system，UMLS)，基因本体(gene ontology，GO)协会开发的 GO 等.

随着生物文献"爆炸式"的增长，非标准的生物名称实体也随之高速增长，在生物领域，这些未经权威认可、自命名的名称使得实体的分类与标注面临着困难[1-2]，而现有的实体识别算法缺乏相应的背景知识，也不能面对生物领域的这种特殊现象.目前，学术界针对这种特殊领域展开了深入的研究，并提出了相应的解决方法，基本归于 3 类，即基于规则[3]、字典[4]和统计学习[5]的方法.2004 年，在文本检索大会(TREC)上基因学项目组明确了该领域文

收稿日期：2007-05-25. 作者简介：崔舒宁(1972—)，男，博士生；冯博琴(联系人)，男，教授，博士生导师. 基金项目：陕西省自然科学基金资助项目(2004F06)；"九八五"二期平台建设资助项目.

献分类和标注的任务.本文通过研究基因生物文献特征,提出了一种能对生物基因文献进行自动标注和分类的方法,并在TREC2004的数据集上对所提算法作了全面的评测.

1 连接MeSH并基于Chi-Square的K最邻近算法

1.1 算法描述

每篇待分类的文献按照逻辑可分为title、abstract、keyword、introduce、method、result、discuss、acknowledge及MeSH 9个逻辑块.设文档D可分为m个逻辑块,其中文档D_i由逻辑块S_{ij}($1\leqslant j\leqslant m$)组成;又设λ_j是第j个逻辑块的权重,λ_{MeSH}是MeSH资源的权重,则文档D_i的向量

$$\boldsymbol{W}_i=\sum_{j=1}^{k}\lambda_j W_{ij}+\lambda_{\mathrm{MeSH}}W_{\mathrm{MeSH}} \tag{1}$$

设T为词表,文档D的向量为$\boldsymbol{W}=(w_1,w_2,\cdots,w_3)$,其中$w_i(1\leqslant i\leqslant p)$为$T$中第$i$个单词项$t$在文档$D$中的权重.若$t$为$T$中的单词项,分类类别为$c_i$,则有[6]

$$\chi^2(t,c_i)=\frac{N(\alpha\nu-\mu\beta)^2}{(\alpha+\mu)(\beta+\nu)(\alpha+\beta)(\mu+\nu)} \tag{2}$$

式中:$\alpha=\|\{D|D\in c_i\wedge t\in D\}\|$;$\beta=\|\{D|D\notin c_i\wedge t\in D\}\|$;$\mu=\|\{D|D\notin c_i\wedge t\notin D\}\|$;$\nu=\|\{D|D\notin c_i\wedge t\notin D\}\|$;$N$是文档的总数.

对于每个t,计算每个类别c_i的$\chi^2(t,c_i)$,选择前n个项作为特征项即可形成逻辑块的n维特征向量.

在上述算法的基础上,本文利用了下述2种分类方法.

(1)K邻近法.设T_{TR}为训练集合,$T_{\mathrm{TR}i}$为训练集中的文档,则训练集中的每一个文档可用n维向量$\boldsymbol{W}$来表示.同样,对待分类的文本D_i,形成向量$\boldsymbol{W}_{D_i}$并计算该向量与T_{TR}中每一篇文档的距离,可得到最近的K个距离.在新文档的K个邻居中,依次计算每类的权重,可得

$$P(\boldsymbol{x},c_i)=\sum_{D_{\notin}K}\mathrm{sim}(\boldsymbol{x},D_i)y(D_i,c_i) \tag{3}$$

式中:$\boldsymbol{x}$为新文档的特征向量;$\mathrm{sim}(\boldsymbol{x},D_i)$为相似度,可通过向量空间模型来计算;$y(D_i,c_i)$为类别属性函数.

文档分配到权重最大类别中的决策规则为

$$y(\boldsymbol{x},c_i)=p(\boldsymbol{x},c_i)-b_i \tag{4}$$

式中:b_i为决策阈值.

(2)最大熵法.设文档类型集合C的个数为l,最初的训练文档集合$T=\{\langle D_i,C_i\rangle|l\leqslant i\leqslant m\}$,其中$m$为训练集合中的文档数目,若每一种逻辑块对应一个新的训练集合$T_{\mathrm{TS}j}=\{\langle S_{ij},C_i\rangle|l\leqslant i\leqslant m\}$,那么在$T_{\mathrm{TS}j}$上利用文本分类法可计算出$S_j$与任意一个类型$C_i$的相似度$\mathrm{sim}_{ij}(C_i,S_j)$.定义文档$D$有$n$个特征$f_i$,则

$$f_i(D,C_i)=\mathrm{sim}_{ij}(C_i,S_j)\qquad 1\leqslant j\leqslant n \tag{5}$$

利用最大熵模型[7],可将文档D分配到类型C_i中概率$P(C_i|D)$为最大的一类中,而

$$P(C_i\mid D)=\frac{1}{Z(D)}\exp\Big(\sum_{j=1}^{n}\lambda_j f_i(D,C_i)\Big) \tag{6}$$

式中:$Z(D)$是归一化因子,即

$$Z(D)=\sum_{k=1}^{l}\exp\Big(\sum_{j=1}^{n}\lambda_j f_i(D,C_i)\Big) \tag{7}$$

1.2 加权法分析

分类算法的关键是,越是重要文本应赋予其特征项的权重就越大.事实证明,采纳χ^2的加权法的分类效果优于单词频率/逆文档频率(TFIDF)加权法.以χ^2值作为横坐标,逆文档频率(IDF)作为纵坐标,可计算文档内容、MeSH特征项的相应位置,其中MeSH特征项的首字母为大写,文档内容特征项的首字母为小写,计算结果如图1所示.

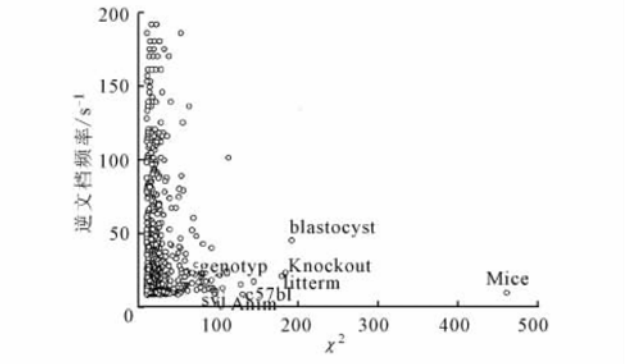

图1 IDF加权和Chi-Square加权的比较

从图中可以看到,具有分类意义的特征项(χ^2值很大)由于被很多文件所包含,所以它的IDF很小.这表明,IDF只是简单地根据包含项的文档数决定该项的分类意义,而χ^2值则从包含项的文档分布来决定项的分类意义.此外,从图1中还可以看到,特征项Mice、Knockout、blastocyst、litterm和Anim(Animmal的词根)等的χ^2值很大,其IDF值却较小,这表明TFIDF加权分类效果不如带有监督的

χ^2加权方案.

2　实验与结果

2.1　实验数据价

实验数据来自 TREC2004 测试数据集，测试集上共 6 043 篇文档，其中有 447 篇属于正例，其余的 5 596篇属于反例. 实验的任务是从测试集中选择出含有老鼠基因的文档，随后再对这些文档进行 GO 标注.

在分类系统中，设 d_T 为正确分类的实际文档数，d_R 为错误分类的实际文档数，d 为实际文档数，U_r 和 U_{nr}分别为正例和反例的评价系数，一般为常数，则在 TREC2004 中的评价规则为

$$U_{norm}=(U_r d_T+U_{nr} d_R)/(U_r d) \tag{8}$$

2.2　分类结果

实验中取 $U_r=20$，$U_{nr}=-1$. TREC2004 的评价标准为 U_{norm}，Triage 任务官方开放测试的最好结果是 $U_{norm}=0.641$. 本文算法的实验结果见表 1.

表 1　逻辑块独立分类的实验结果

	U_{norm}								
	标题	摘要	关键字	导言	方法	结果	讨论	感谢	MeSH
开放测试	0.374	0.495	0.374	0.426	0.485	0.544	0.432	0.374	0.661
封闭测试	0.467	0.680	0.370	0.622	0.635	0.645	0.610	0.449	0.714

本文算法的测试集分类结果与 TREC 公布结果的比较如表 2 所示，其中邻居数 $K=63$. 从中看出，正例类的阀值能使 U_{norm} 值达到最优，其中加权法 TFCHI1 的分类效果已经达到和超过了 TREC 官方的最好效果. 将每一个逻辑块的向量直接与 MeSH 的向量连接在一起，然后用连接后的向量进行分类，结果如表 3 所示. 选择摘要、导言、方法、结果，讨论以及 MeSH 6 个逻辑块作为最大熵模型的特征，训练用最大熵模型权重见表4，根据表 4 的权重进行分类测试可得到开放测试下的 $U_{norm}=0.543$，封闭测试下的 $U_{norm}=0.846$.

2.3　标注结果

要利用分类器完成基因本体标注，首先要进行基因的文本碎片抽取，提取范围称为粒度. 本文的粒度内容有：整篇文档(Article)，段落(Paragraph)，包含基因的一句话(G)，包含基因的一句话和前一句话(P+G)，包含基因的一句话和后一句话(G+S)，包含基因的一句话和前、后各一句话(P+G+S)，基因名(Gene). 在 TREC2004 标注任务中，开放测试公布的最佳 U_{norm} 值为 0.561，而每一个粒度独立实验的结果如表 5 所示，算法的每一个粒度连接 MeSH 的实验结果如表 6 所示.

表 2　本文算法分类结果与 TREC 结果之比较

		查全率	查准率	U_{norm}
TREC	Best	0.153	0.888	0.641
	Worst	0.200	0.014	0.011
	Mean	0.183	0.519	0.330
	TFIDF	0.092	0.765	0.386
本文算法	TFCHI1	0.151	0.935	0.672
	TFCHI2	0.141	0.886	0.617

表 3　连接 MeSH 词汇表算法的分类结果

逻辑块测试	U_{norm}							
	标题＋MeSH	摘要＋MeSH	关键字＋MeSH	导言＋MeSH	方法＋MeSH	结果＋MeSH	讨论＋MeSH	感谢＋MeSH
开放测试	0.657	0.655	0.656	0.659	0.630	0.660	0.633	0.652
封闭测试	0.728	0.750	0.714	0.754	0.718	0.709	0.747	0.729

表4 摘要、导言、方法、结果、讨论以及MeSH在最大熵算法中的权重系数

	摘要	导言	方法	结果	讨论	MeSH
MeSH权重系数	0.097	0.089	0.068	0.080	0.073	0.383

表5 每一个粒度的独立实验结果

	Article	Paragraph	G	P+G	G+S	P+G+S	Gene	MeSH
开放测试粒度	0.537	0.538	0.537	0.556	0.554	0.552	0.335	0.420

表6 每一个粒度连接MeSH的实验结果

	Article+MeSH	Paragraph+MeSH	G+MeSH	P+G+MeSH	G+S+MeSH	P+G+S+MeSH	Gene+MeSH
粒度	0.549	0.569	0.542	0.569	0.564	0.584	0.394

3 结 论

从分类的实验结果可以看到:①文档算法的每一个逻辑块对文档分类的贡献程度不一样,对于基因的本体标注,MeSH的贡献最大;②采用向量连接的方法和最大熵模型组合方法,可以提高封闭测试的分类效果,但并非能够提高开放测试的分类效果;③逻辑块独立分类实验的最佳$U_{norm}=0.661$,比官方测试的最好结果提高了3.12%.

从标注的实验结果可以看到:①经基因本体标注后,对包含基因的语句及其前、后语句的提取效果最好;②采用连接MeSH向量的方法,可普遍提高开放测试分类的效果;③粒度法标注实验的最好结果是0.584,比官方测试的最好结果提高了4.10%.

参考文献:

[1] DIETRICH REBHOLZ-SCHUHMANN H K, COUTO F. Facts from text — is text mining ready to deliver? [J]. PLoS Biology, 2005, 3(2):188-191.

[2] ADITYA V P B, KINCAID R. An architecture for biological information extraction and representation [J]. Bioinformatics, 2005, 21(4):430-438.

[3] NARAYANASWAMY M, RAVIKUMAR K E. A biological named entity recognizer [C]// Proceedings of Pacific Symposium on Biocomputing. Hawaii, USA: World Scientific, 2003:427-438.

[4] TSUJI J. Boosting precision and recall of dictionary-based protein name recognition [C]// Proceedings of Atomic Level Characterizations. Hawaii, USA: Wiley Publisher, 2003:41-48.

[5] ZHOU Guodong, SU Jian. Named entity recognition using an HMM-based chunk tagger [C]// Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. San Francisco, USA: Morgan Kaufmann Publishers, 2002:473-480.

[6] YANG Yiming, PEDERSEN J O. A comparative study on feature selection in text categorization [C]// Proceedings of the 14th International Conference on Machine Learning. San Francisco, USA: Morgan Kaufmann Publishers, 1997:412-420.

[7] ROSENFELD R. A maximum entropy approach to adaptive statistical language modeling [J]. Computer, Speech, and Language, 1996(10):187-228.

(编辑 苗凌)