基于语义关联和信息增益的 TFIDF 改进算法研究*

许 珂, 蒙祖强, 林啓峰
(广西大学 计算机与电子信息学院, 南宁 530004)

摘 要: 基于词频反文档频率(term frequency inverse document frequency,TFIDF)的现有文本特征提取算法及其改进算法未能考虑类别内部词语之间的语义关联,如果脱离语义,提取出的特征不能很好地刻画文档的内容。为准确提取特征,在信息熵与信息增益的基础上,加入词语的语义关联因素,实现融合语义信息的特征提取,进而提出语义和信息增益相结合的 TFIDF 改进算法,该算法弥补了统计方法丢失语义信息的弊端。实验结果表明,该算法有效地提高了文本分类的精准率。

关键词: 词频反文档频率; 特征提取; 语义关联; 信息增益; 文本分类
中图分类号: TP301 **文献标志码**: A **文章编号**: 1001-3695(2012)02-0557-04
doi:10.3969/j.issn.1001-3695.2012.02.041

Improved TFIDF feature extraction algorithm based on semantic association and information gain

XU Ke, MENG Zu-qiang, LIN Qi-feng
(*College of Computer & Electronic Information, Guangxi University, Nanning* 530004, *China*)

Abstract: Both the traditional and improved term frequency-inverse document frequency (TFIDF) algorithms ignored the difference of distributions among different categories in feature extraction. Due to the lacking of consideration of semantic relationships within some certain categories, the selected feature word cannot describe the contents of the document correctly and accurately. In order to select feature more accurately, in this paper, based on the previous improvements, introduced the semantic association of words to analyze the semantic of text, redesigned the weights equation, and proposed the new TFIDF algorithm combined with semantic and information gain. The developed algorithm can make up for the shortcomings of the lack of semantic information in statistical method. Experimental results illustrate that the improved algorithm can effectively improve text classification accuracy.

Key words: TFIDF; feature extraction; semantic association; information gain; text classification

0 引言

随着网络技术的飞速发展,互联网上的电子文本数量急剧增加,人工分类已经无法满足要求,因此如何有效地进行文本分类及管理是当今的一个重要研究方向,也为文本分类提供了新的发展平台[1]。文本分类通常有文本的预处理、文本的向量空间模型表示、文本特征提取和分类器的训练四个步骤。在研究文本分类的过程中,特征提取是最关键的环节之一,可以起到降低向量维数、简化计算、去除噪声等作用。故而,特征提取的好坏将直接影响文本分类的准确率。特征提取的基本思想是构造一个评估函数,对特征集中的每个特征词进行权重计算,然后对所有的特征词按照其权重值的大小进行排序,选择预定数目的最佳特征作为最终的特征子集。在文本分类中常使用的特征选择方法包括文档频率(document frequency)、互信息(mutual information)、信息增益(information gain)、χ^2 统计(CHI)、期望交叉熵(cross entropy)、文本证据权(the weight of evidence for text)、优势率(odds ratio)和词频反文档频率等[2]。对于传统的 TFIDF 特征提取算法,目前已有文献[36]对其的不足进行分析及改进,效果比较显著的是通过引入信息熵对其进行改进,解决了词语在类别间的分布不均造成的问题。比如字面不同的词语但可以表示同一个含义,这样的一组词的语义特征[710]是一样的,不能忽视它们共同出现对词频的影响。若忽视了语义,就无法准确表达文档的内容,同时也会影响计算特征词权值的精确度。之前的大多工作是计算独立的字或词的权重值,以选出特征词,很少涉及词汇语义部分的研究,容易忽视多义词和同义词现象。本文针对这一问题,先对词语进行语义信息的分析,然后将有语义关联的归为一组,进而分析信息熵,改进信息增益的公式,提出一种新的基于语义关联和信息增益的 TFIDF 特征选择算法。实验结果表明,改进后的特征选择算法,在文本分类的查准率和召回率两方面均有不同程度的提高。

1 相关工作

1.1 TFIDF 特征选择

在文本分类领域中,最常用的是 Salton 在 1975 年提出的向量空间模型(VSM)。VSM 将文本 d_i 看做向量空间中的一个

收稿日期: 2011-07-24; **修回日期**: 2011-08-25 **基金项目**: 国家自然科学基金资助项目(61063032); 广西教育厅科研基金资助项目(201012MS010)

作者简介: 许珂(1987-),女,河南新乡人,硕士研究生,主要研究方向为智能系统与智能 CAD(xukeability@126.com);蒙祖强(1974-),男,广西罗城人,教授,博士,主要研究方向为人工智能、数据挖掘;林啓峰(1987-),男,福建三明人,硕士研究生,主要研究方向为数据挖掘.

n 维向量$(t_1, w(t_1), t_2, w(t_2), t_3, w(t_3), \cdots, t_n, w(t_n))$，则 $t_1, t_2, t_3, \cdots, t_n$ 是该文本的特征词，$w(t_i), i = 1, 2, 3, \cdots, n$ 是该文本对应的第 i 个特征词的权重值。

对文本文档进行分类主要依据文档的内容，而特征词的权重值便是刻画词语表达文档内容的重要指标。权重值的计算按其值类型通常分为以下两种[11]：a)布尔型，即将所有训练文档的词语作为全集，当一个词语 t_i 出现在文档中时，其权值设为 1，否则设为 0；b)实数型，将文档的词语通过权重计算公式求出其权重值。TFIDF 是 VSM 中经典的特征权值函数，权重计算公式为

$$\text{weight}_{\text{tfidf}(t)} = \text{tf}(t) \times \text{idf}(t)$$

其中：tf(term frequency)为词语频率，表示该词语在文档中出现的次数；idf(inverse document frequency)为反文档频率，表示该词语在文档集合中分布情况的量化。通常计算 idf 的方法为

$$\text{idf}(t) = \log \frac{N}{n}$$

其中：N 为文档集中的总文档数，n 为出现特征项 t 的文档数。

显然，传统的 TFIDF 特征选择方法中，某个词语的权重值与该词语的频率成正比，与文档频率成反比。但这个方法有着明显的不足，即忽视了文档在每个类中的分布情况。对于文档频率，一方面只考虑了包含某个词语文档数绝对量的多少，而没有考虑这些文档在类别中的分布；另一方面，假如说包含某词条的文档数比较少，但如果这个词语均匀分布于各个类别中，那么对分类的贡献是微乎其微的，不能很好地区分类别。相应地，它的权重值应该比较小，但是按照传统 TFIDF 算法得出的权重值却比较大。上述两个明显的缺点主要是因为 TFIDF 将文档集合作为整体考虑，没有考虑词语在类别间的分布情况。针对这个问题，文献[10]对传统 TFIDF 方法进行了改进，引入了信息熵与信息增益的概念，用以解决词语在类别间的分布不均。但文献[10]在处理方法上未考虑同一个文档中词与词之间的语义关联，只是将每个词语孤立地进行权重值的计算，这样的处理将词语割裂开，不利于文本内容表达的完整性，对文本分类有一定的影响。本文在基于语义关联的前提下计算信息熵，对权重值的计算方法进行改进。

1.2　基于信息熵的特征选择

熵是德国物理学家克劳修斯于 1850 年提出的，表示一种能量在空间中分布的均匀程度，能量分布得越均匀，熵就越大。1948 年，Shannon 把熵应用于信息处理，提出了信息熵的概念。信息熵在随机事件发生之前，是结果不确定性的量度；在随机事件发生之后，它是人们从该事件中所得到信息的量度(信息量)。

设随机事件 X 在获得信息 y 之前结果的不确定性为 $H(X)$，得到信息 y 之后为 $H(X|y)$，那么包含在消息 y 中的关于事件 X 的信息量为

$$G(X, y) = H(X) - H(X|y) \tag{1}$$

条件熵 $E(X|y) = H(X|y)$ 是观测信息 y 后信息空间 X 的不确定程度。信息增益是信息熵的差，表示为

$$\text{IG}(X, y) = H(X) - H(X|y) \tag{2}$$

由公式得出的不确定程度减少量就是信息增益，即表示词语 y 对分类的影响。倘若简单地将信息增益作为一个乘数因子加入 TFIDF 中，修改 TFIDF 算法中的权重公式为 tf × idf × IG，并不能解决传统 TFIDF 的不足，所以在 tf × idf × IG 公式的基础上，将信息增益公式进行变形并引入到文档集合的类别间，将文档类别看做信息源，由训练数据集合的类别信息熵和文档类别中词语的条件熵之间信息量的增益关系共同决定该词语在文本分类中所提供的信息量，即建立起信息熵和词语权重值之间的关系。则权重值的计算公式为

$$\text{weight}_{\text{tfidf}}(t_i) = \text{tf} \times \text{idf} \times \text{IG}(C, t_i) \tag{3}$$

其中：

$$\text{IG}(C, t_i) = E(C) - E(C/t_i) \tag{4}$$

$$E(C) = -\sum_{j=1}^{m} p(C_j) \times \log(p(C_j)) \tag{5}$$

$$E(C/t_i) = -\sum_{j=1}^{m} p(C_j/t_i) \times \log(p(C_j/t_i)) \tag{6}$$

对式(3)研究得出，可以解决传统 TFIDF 中存在的不足，即当词语 t 在类别中分布不均匀时，在某个类别中大量出现而其他类别中分布较少，理论上这个词带有很大的类别信息，由改进后的公式也恰恰算出它的权重值较高。另一种情况是某个词语虽然在整个文档集合中数量很少，但均匀分布于各个类别间，则其对区分类别的影响比较小，理论上它的权重值相应地比较低，由式(3)算出的权重也确实比较低。

2　基于语义关联的改进特征选择方法

式(3)虽然弥补了词语在类别间分布上的不足，但针对类别内部，式(3)却未考虑词与词之间的语义关联。当一个词语在类别内部经常重复出现，但与这个词有语义关系的词语却非常稀少，那么这个词与文档所描述的内容关联不大，不能有效地用来分类，应该赋予小的权重值，但其出现次数多，TF 值较大，则算出的权重值反而偏大。同时另外一些词，虽然单个出现的次数并不多，但它们相互之间关系非常密切，几乎同时出现与共存，它们在文档中的同时出现对文档内容的描述有着重要影响。常见的在文档中存在这样三类词：字面不同但表示的含义相同，一词多义，以及上下文之间隐含的语义关系。这种现象产生的原因便是词汇层面(代表意思的词)和概念层面(意思本身)的差别，在文档中这些词的这种关系体现为语义关联，这样会导致特征词提取的不准确。解决办法是可以通过计算语义相似度[12]将这些存在语义关联的词归为一组，这样便利用语义特征将这些词联系起来。将它们划归为一组后，再一同计算信息增益值，最终对文档分类提供的信息量就会变大。

结合语义关联，本文提出对文本文档中类别内部的词语先进行分词，进而合并有语义关联的词语，将这些词划归一组。在 $\text{tf} \times \text{idf} \times \text{IG}_e$ 中，计算 IG_e 中的条件熵部分时，不以单个词为出发点，而考虑某类别内部出现的从属于同一组的词语，先计算这组词在某个类别中出现的概率和，然后计算这组词的条件熵，最后求信息增益。即

输入：用于训练的文档集 D。

输出：融合语义的 KNN 分类器。

step 1. 选用中科院的 ICTCLAS 系统对文本进行分词，词语表示为 $t_1, t_2, \cdots, t_n$；

step 2. 同一类别内部的词语进行归并成组；

i = 1；

while ($t_i \in (t_1, t_2, \cdots, t_n)$)

{ switch (依据知网[13]的各种关系定位 t_i，判断属于哪一组)

case 1；

　则 t_i 归属关系组 1；

case 2；

　则 t_i 归属关系组 2；

……

case k；

　则 t_i 归属关系组 k；

```
default;
i + + ;
}
```

step 3. 对同一个类别内部,计算被归为一组的词语出现的概率和 $\sum_{i=1}^{n} p(C_j/t_i)$;

```
while(t_i ∈ (t_1, t_2, ⋯, t_n))
{
  while(t_j ∈ (t_1, t_2, ⋯, t_n) && j≠i)
  {
  if (t_i 与 t_j 基于知网存在语义上的 某种关联)
    p(C_j/t_j) + = p(c_j/t_j);
else
    输出 p(c_j/t_i);
    j + + ;
  }
i + + ;
}
```

然后分别求每个词的条件熵 $E(C/t_i)$ 以及信息增益 $\overline{IG_e}(C,t) = E(C) - \overline{E(C/t_i)}$;

step 4. 计算每个词语的权重值 $weight_{tfidf}(t_i)$;

step 5. 对文档中词语的权重值 $weight_{tfidf}(t_i)$ 进行排序,选出权重值较大的前 N 个作为特征词;

step 6. 采用 KNN 算法对训练文档进行分类。

本文针对步骤 2、3 进行深入研究。步骤 2 中找出有语义关联的词语进行合并时所依据的是知网[13]中的各种关系。在知网中,每个词可以具有多个概念(一词多义),此外知网还定义了上下位关系、同义关系、反义关系、部件—整体关系等 16 种关系,通过一系列的义原,利用某种知识描述语言来描述一个概念。将词特征和概念特征联系起来,通过索引表可以建立概念与关系之间的联系,那么对于任给的一个词,可以对其定位,快速查找到与之有关联的其他词语,然后将这些词归入一组。

综上,归组完毕后需要修改条件熵 $E(C/t_i)$ 为

$$\overline{E(C/t_i)} = -\sum_{j=1}^{m}\left[\left(\sum_{i=1}^{n} p(C_j/t_i)\right) \times \log\left(\sum_{i=1}^{n} p(C_j/t_i)\right)\right]$$

其中:C 为文档集合,m 为文档类别的个数,n 为词语 t_i 所在的分组中词语的总个数,$p(C_j/t_i)$ 表示词语 t_i 在类别 C_j 中出现的概率。

则信息增益公式修改为

$$\overline{IG_e}(C,t) = E(C) - \overline{E(C/t_i)}$$

最终的权重公式为

$$weight_{tfidf}(t_i) = TF \times IDF \times \overline{IG_e(C,t)}$$

利用本文改进后的条件熵及信息增益公式完成步骤 3 的计算。

改进后的权重公式在保证很好地反映词语在类别间的分布情况基础上,将类别内部词语的语义关联加入公式中,对文档分类提供更全面的信息量,解决了传统 TFIDF 算法的不足,修正了权重值的计算公式,继而便可进行特征提取,完成文本分类的验证实验。实验流程如图 1 所示。

3　实验分析

为了验证本文对信息增益公式的改进可以提高分类的准确度,在搜狗网站上下载了 1 000 篇文档作为语料,用于文本分类的实验,主要有教育、体育、财经、工业、医药、政治、汽车、军事、艺术、计算机共十个方面的内容。取训练文本 500 篇,测试文档 500 篇。采用中科院的 ICTCLAS 系统(http://ictclas.org/)对所有文本进行分词,用 C + + 语言实现文本分类。实验采用 $tf \times idf \times IG_e$ 和本文提出的 $tf \times idf \times \overline{IG_e}$ 公式来计算词语的权重值,然后采用 KNN 算法[14]进行分类,分别从查准率和召回率两个评价指标对两种权重计算方法的分类效果进行比较。查准率和召回率是两个常用的文本分类评估测试值,查准率是指测试集中被正确分类的文档数除以测试集中被分类为正例的文档数;召回率等于正确归类的文档数与应有的文档数的比值。在 KNN 分类算法中,选取 $K = 14$ 时分类的实验结果分别如表 1、2 所示。

图1　文本分类流程

表 1　基于信息增益的 TFIDF 算法的分类结果

分类	教育	体育	财经	工业	医药	就业	汽车	军事	艺术	计算机
教育	47	0	0	1	0	1	0	1	0	0
体育	0	46	1	0	0	1	1	0	1	0
财经	0	0	47	0	0	1	0	0	1	1
工业	1	1	1	42	1	1	1	1	0	1
医药	0	0	1	0	46	1	0	1	0	1
就业	0	1	1	1	0	46	1	0	0	0
汽车	1	1	1	0	0	1	45	1	0	0
军事	1	1	1	1	1	1	1	41	1	1
艺术	0	1	1	1	0	1	0	1	45	0
计算机	0	1	1	1	1	1	1	1	0	43
查准率/%	94.00	89.03	85.40	90.83	93.55	83.78	92.71	87.44	96.40	91.48
召回率/%	94.00	92.00	94.00	84.00	92.00	92.00	90.00	82.00	90.00	86.00

其中:每个类别为 50 个样本,平均查准率为 90.57%,平均召回率为 89.60%。

表 2　基于语义关联和信息增益的 TFIDF 改进算法的分类结果

分类	教育	体育	财经	工业	医药	就业	汽车	军事	艺术	计算机
教育	49	0	0	0	0	0	0	0	1	0
体育	0	48	0	1	0	1	0	0	0	0
财经	0	0	47	0	0	1	0	0	1	1
工业	0	0	1	48	0	0	0	0	0	1
医药	1	0	0	1	45	0	1	1	1	0
就业	0	0	0	1	0	49	0	0	0	0
汽车	0	0	0	1	0	0	48	1	0	0
军事	0	1	0	0	1	1	1	46	0	0
艺术	1	1	1	0	1	1	1	1	42	1
计算机	0	1	1	0	1	0	0	0	1	46
查准率/%	96.08	94.12	94.00	92.31	93.75	92.45	94.12	93.88	91.30	93.88
召回率/%	98.00	96.00	94.00	96.00	90.00	98.00	96.00	92.00	84.00	92.00

其中:每个类别为 50 个样本,平均查准率为 93.59%,平均召回率为 93.60%。

对实验结果分析得出,使用改进后的权重公式,查准率和召回率均有不同程度提高,在平均查准率和平均召回率两个指标中提高体现得更明显。实验表明在类别内部词语之间的语义关联对权重值有不可忽视的影响,将这个因素引入公式中对最终的分类效果有明显的提高。

为了进一步验证考虑语义关联后,文本的分类效果会有显著的提高,以计算机和艺术类别的文本为例设计如下实验。选用上述实验测试集中的计算机类文档 50 篇,艺术类文档 50 篇,继续采用中科院的 ICTCLAS 系统(http://ictclas.org/)对所有文本进行分词,用 C + + 语言实现文本分类。

此实验中对语料进行了特殊处理,比如增加文档中同义词的比例。在原始语料中,依托知网中的词语关系,通过人工统

计,计算机类文档中有语义关联的词语占总词数的平均比例为14.9%,艺术类文档中有语义关联的词语占总词数的平均比例为10.2%。为凸显本文加入语义关联后对特征提取的效果,进行强语义关联下的文本分类实验。人为地对两个类别的文档分别增加同义词,使得计算机类文档中有语义关联的词语比例提高到30%,艺术类文档中有语义关联的词语提高到20%,这样做的目的使得类别更鲜明,更有利于进行数据分析,针对专业性较强的文本有明显的优势。依然采用本文提出的 $tf \times idf \times \overline{IG}_c$ 进行权重值的计算,进而用 KNN 算法进行文本分类。分类效果如表3所示。

表3 强语义关联下的分类效果比较

分类	原始语料		处理后的语料	
	计算机	艺术	计算机	艺术
计算机	45	8	49	3
艺术	5	42	1	47
查准率/%	84.91	89.36	94.23	97.80
召回率/%	90.00	84.00	98.00	94.00

由表3中数据得知,计算机类的查准率在进行强语义关联处理后查准率提高了9.32%,召回率提高了8%;艺术类的查准率提高了8.44%,召回率提高了10%。由此可见,在强语义关联下,本文的方法优势很明显。中文信息中有大量语义关联的词汇,所以本文的方法有较大的应用价值,对文本分类的效果的提高有显著帮助。

4 结束语

本文对基于信息增益的 TFIDF 算法进行了深入研究,将文档类别看做信息源,在考虑词条在文档各类别间的分布情况的基础上,加入词与词之间的语义关联这一重要因素重新计算信息增益,从而提高了文本分类的准确度。因为特征选择算法大多对文本分类算法存在一定的依赖关系,所以在未来的工作中,将本文提出的改进方法与 Native Bayes 等文本分类方法相结合,在期望得到更好的分类效果的同时检验该特征算法对分类算法的依赖情况。

参考文献:

[1] SEBASTIAN F. Machine learning in automated text categorization [J]. ACM Computing Survey,2002,34(1):1-47.

[2] 胡佳妮,徐蔚然,郭军,等. 中文文本分类中的特征选择算法研究[J]. 光通信研究,2005(3):44-46.

[3] 张瑜,张德贤. 一种改进的特征权重算法[J]. 计算机工程,2011,37(5):211-212.

[4] 周炎涛,唐剑波,王家琴. 基于信息熵的改进 TFIDF 特征选择算法[J]. 计算机工程与应用,2007,43(35):156-158.

[5] 陈瑞芬. 一种结合反馈方法的中文文本分类算法[J]. 计算机应用,2005,25(12):2862-2864.

[6] 罗杰,陈力,夏德麟,等. 基于新的关键词提取方法的快速文本分类系统[J]. 计算机应用研究,2006,23(4):32-34.

[7] 方俊,郭雷,王晓东. 基于语义的关键词提取算法[J]. 计算机科学,2008,35(6):148-150.

[8] 马军,宋玲,韩晓晖,等. 基于网页上下文的 Deep Web 数据库分类[J]. 软件学报,2008,19(2):267-274.

[9] 施洋,张奇,黄萱菁. 含有语义特征的网页新闻自动抽取[J]. 计算机工程,2010,36(7):173-178.

[10] 许莉,王大玲,夏秀峰. 基于句法和语义信息的问句特征提取方法[J]. 计算机工程,2010,36(21):65-70.

[11] 张玉芳,陈小莉,熊忠阳. 基于信息增益的特征词权重调整算法研究[J]. 计算机工程与应用,2007,43(35):159-161.

[12] 陈悦,张蕾,胡艳华. 基于语义的信息检索模型[J]. 计算机工程与应用,2009,45(26):141-143.

[13] 董振东,董强. 知网简介[EB/OL]. http://www.keenage.com.

[14] 王煜. 基于决策树和 K 最近邻算法的文本分类研究[D]. 天津:天津大学,2006.

(上接第545页)性。与其他方法相比,矩阵法检测同源性有其自身的优越性,特别是在多序列的逐步搜索过程中。在以后的研究中,笔者将进一步探讨矩阵理论在生物信息学其他方面的应用。

参考文献:

[1] DESESQUELLES P, HA T M H, HAUSCHILD K, *et al*. Matrix formalism and singular-value decomposition for the location of gamma interactions in segmented HPGe detectors[J]. The European Physical Journal A: Hadrons and Nuclei,2009,40(2):237-248.

[2] SCHWAENENA C, NESSLINGA M, RADLWIMMERA B, *et al*. Application of matrix-CGH (array-CGH) for genomic research and clinical diagnostics[C]//Microarray Technology and Its Application. Berlin: Springer,2005:251-263.

[3] NAGARAJU A, RAMACHANDRAM S. Ad-hoc on demand distance vector routing algorithm using neighbor matrix method in static Ad-hoc networks[C]//Advances in Networds and Communications. Berlin: Springer,2011:44-54.

[4] KOWALUK M, LINGAS A. Unique lowest common ancestors in dags are almost as easy as matrix multiplication[C]//Proc of the 15th Annual European Conference on Algorithms. Berlin: Springer. 2007: 265-274.

[5] OWEN R. Lectures on the comparative anatomy and physiology of the invertebrate animals[M]. Lodon: Longman, Brown, Green & Longmans,2010:122-193.

[6] 丁彦蕊,蔡宇杰,孙俊,等. 用同源建模法探索提高脂肪酶的耐热性[J]. 计算机与应用化学,2007,24(11):1535-1538.

[7] SHAN Hong-yan. Morphology, Molecules and homology[J]. Chinese Bulletin of Botany,2007,24(1):71-79.

[8] TAMURA K, DUDLEY J, NEI M, *et al*. MEGA4: molecular evolutionary genetics analysis (MEGA) software version 4.0[J]. Molecular Biology Evolution,2007,24(8):1596-1599.

[9] 刘泽隆,于发科. Vector NTI Suite 在分子生物学领域的应用[J]. 生物学杂志,2006,23(4):13.

[10] ALTSCHUL S F, MADDEN T L, SCHAFFER A A, *et al*. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs[J]. Nucleic Acids Research, 1997, 25(17): 3389-3402.

[11] ROCIO G D, JIMENEZ M J, MEDRANO M, *et al*. Integral operators for computing homology generators at any dimension[C]//Proc of the 13th Iberoamerican Congress on Pattern Recognition: Progress in Pattern Recognition, Image Analysis and Applications. 2008:356-363.

[12] NIYOGI P, SMALE S, WEINBERGER S. Finding the homology of submanifolds with high confidence from random samples[J]. Discrete & Computational Geometry,2008,39:419-441.

[13] SHARP P M, LI W H. An evolutionary perspective on synonymous codon usage in unicellular organisms[J]. Journal of Molecular Evolution,1986,24(1-2):28-38.

[14] CHIAPELLO H, LISACEK F, CABOCHE M, *et al*. Codon usage and gene function are related in sequences of Arabidopsis Thaliana[J]. Gene,1998,209(1):1-38.