基于非对称多值特征杰卡德系数的高维语义向量差异性度量方法

冯艳红[1,2] 于 红[1,2] 孙 庚[1,2] 彭 松[1]
(大连海洋大学信息工程学院 大连 116023)[1]
(大连海洋大学辽宁省海洋信息技术重点实验室 大连 116023)[2]

摘 要 语义向量差异性度量是采用深度学习方法解决自然语言处理领域问题的重要基础。在高维语义向量差异性度量中存在“度量集中”问题,导致通过传统的度量方法得到的度量结果无法体现语义向量间的差异性。针对该问题,提出一种基于非对称多值特征杰卡德系数的差异性度量方法。由高维语义向量维度值的统计分布得出,部分维度的维度值密集地分布在特定值域内,导致其无法贡献差异度,因此不同维度对差异性的贡献量不同,具有非对称性。该方法定义了关于维度值的重要性函数,选取重要性函数值满足阈值的维度参与差异度计算,去掉无法贡献差异度的维度,从而实现了降维,缓解了“度量集中”问题。分别在渔业数据集和公开数据集上,对不同维度的语义向量的不同度量方法进行了比较,结果表明在语义性没有明显变差的情况下,所提方法的多样性指标较目前最优的度量方法有大幅提高。

关键词 非对称多值特征,杰卡德系数,高维语义向量,度量方法,度量集中
中图法分类号 TP183 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2018.06.010

Diversity Measures Method in High-dimensional Semantic Vector Based on Asymmetric Multi-valued Feature Jaccard Coefficient

FENG Yan-hong[1,2] YU Hong[1,2] SUN Geng[1,2] PENG Song[1]
(College of Information Engineering, Dalian Ocean University, Dalian 116023, China)[1]
(Key Laboratory of Marine Information Technology of Liaoning Province, Dalian Ocean University, Dalian 116023, China)[2]

Abstract The diversity measures of semantic vector are important base of natural language processing problem resolved by deep learning methods. There is a problem of “measurement concentration” in the diversity measure of high dimension semantic vector, which leads to the diversity of the semantic vectors disappear when the diversity are obtained by the traditional measure methods. To resolve this problem, a diversity measures method based on the asymmetric multi-valued feature Jaccard coefficient was proposed. From the statistical distribution of the dimension values of the high-dimensional semantic vector, the values of the partial dimensions are densely distributed in a certain range, which makes them impossible to contribute the diversity. Therefore, the contribution of different dimensions to the diversity is different and has asymmetry. This method defines the importance function about the dimension value, selects the dimensions of the importance function value satisfying the threshold to participate in the diversity calculation and removes the dimensions that can not contribute the diversity, and then realizes the dimensionality reduction and alleviates the problem of “measurement concentration”. The experiments were respectively conducted on fishery data sets and public data sets. Different measures methods of the different dimension semantic vector were compared. Under the condition that the semantic nature is not markedly reduced, the diversity index of the proposed method is much higher than the current optimal measures method.

Keywords Asymmetric multi-valued feature, Jaccard coefficient, High-dimensional semantic vector, Measures method, Measurement concentration

1 引言

在解决自然语言处理领域的命名实体识别、情感分析、信息检索、问答系统、机器翻译等任务时,通常将文本拆分为词语序列。为方便利用深度学习方法完成这些任务,需要将词语向量化为词向量[1]。本文研究的词向量是利用语言模型训

本文受大连市科技计划项目:海洋渔业大数据管理与集成关键技术研究(2015A11GX022),辽宁省大学生创新创业项目:渔业领域智能问答系统的研究与实现(201710158000131)资助。

冯艳红(1980-),女,硕士,讲师,CCF会员,主要研究方向为自然语言处理、机器学习;于 红(1968-),女,博士,教授,主要研究方向为数据挖掘、信息检索,E-mail:yuhong@dlou.edu.cn(通信作者);孙 庚(1979-),男,硕士,副教授,主要研究方向为嵌入式系统;彭 松(1993-),男,硕士生,主要研究方向为自然语言处理、机器学习。

练深度神经网络得到的，这种词向量具有一定的语义含义[2-3]，故本文称其为语义向量。文本中全部词语的语义向量构成语义向量集合。在自然语言处理的诸多任务中都需要度量该集合中语义向量之间的差异性或相似性，例如，文献[4]为比较词语与领域术语的语义差异性，计算了语义向量的余弦距离，改善了命名实体的识别效果。在推荐系统任务中，文献[5]使用余弦函数和杰卡德相似系数计算了用户和歌曲之间在流派、上下文和情感空间上的相似度。当语义向量的维度不高时，其差异性度量结果比较准确，改善了应用的效果。但利用语言模型训练深度神经网络得到语义向量时[6-7]，为使其蕴含语料中更多有价值的信息，具有更强的表达能力，通常设置较高的维度。根据文献[8]给出高维数据的分析，在文献[6-7]实现的语言模型中，设语言模型的参数为 m_p 个，训练样本为 m_e 个，由于实际应用中训练样本通常不足，导致 m_p 和 m_e 满足 $\log(\frac{m_p}{m_e})\geqslant 1$，此时得到的语义向量为高维语义向量。随着向量维度的增加，为了达到同样的效果，训练模型所需要的样本数将呈指数增长[9]。而实际应用中样本的数量受限，出现了“小样本”问题[10]。从数学分析角度，将“小样本”问题描述为：用传统的欧氏距离表达高维数据的距离时，数据间的距离几乎相同，出现了“度量集中”问题[11]，数据间的差异性无法得到体现，导致传统的差异性度量方法在深度学习中基本失效。

对于“小样本”导致的“度量集中”问题，学者们给出了初步的解决方法。文献[12]用基于网格的划分法将高维空间划分为低维子空间，用子空间上数据的距离近似表达高维空间上数据的距离。理论分析表明，该方法可解决高维数据的“度量集中”问题，但其有效性没有通过实验验证。文献[13]从几何学和统计学的角度给出了“度量集中”现象的直观解释，分析并比较了基于 L_p 范数的度量方法，得出利用分数范数（$p<1$）的度量方法可缓解“度量集中”问题的结论，并通过实验验证了该结论；该文献还证明了单位超立方体的体积主要集中在超立方体的外壳上，即高维数据的某个度量实际上分布在某个维度较低的子空间，这是用降维解决“度量集中”问题的重要理论依据之一。降维在解决高维数据的其他问题方面已取得了良好的效果[14-15]。根据以上文献可得出两点结论：1）降维是一种解决“度量集中”问题的有效方法；2）文献[13]提出的分数范数度量方法及文献[16]中的其他度量方法都是将语义向量看作连续的实数向量。理论上，离散数据较连续数据更能加大数据之间的距离，进而增加多样性，缓解“度量集中”问题。杰卡德系数是一种有效的度量二值离散数据的差异性的方法[17]，其中非对称二值特征杰卡德系数仅将值 1 作为度量差异性的特征，取得了更优的差异性度量结果[18]。本文以此方法为基础，提出一种基于非对称多值特征杰卡德系数的差异性度量方法。该方法利用降维的思想为语义向量的不同维度值赋予不同的重要性，根据重要性函数值确定参与差异性计算的维度，缓解了“度量集中”问题。

本文第 2 节从统计分析的角度阐述了“度量集中”问题产生的原因。由于本文提出的方法基于离散数据，因此第 3 节首先阐述了语义向量的离散化方法，然后在非对称二值特征杰卡德系数的基础上提出一种基于非对称多值特征杰卡德系数的差异性度量方法，以解决“度量集中”问题。第 4 节通过实验将本文提出的方法与语义向量的多种度量方法进行比较，验证了本文提出的度量方法的度量效果大幅优于已有的度量方法。最后总结全文。

2 “度量集中”问题的分析

由于度量方法中的差异度来自语义向量的各个维度间的差异度，因此本文主要对语义向量的维度进行统计分析，给出了度量方法的“度量集中”问题的直观解释。

设文本 $\boldsymbol{w}$ 可拆分为词语序列 $(w_1,\cdots,w_t,\cdots,w_T)$，词语 w_t 的语义向量 $\boldsymbol{x}=(x_1,\cdots,x_j,\cdots,x_d)$。设 $(w_1,\cdots,w_t,\cdots,w_T)$ 生成的词表记为 $\boldsymbol{D}$，$\boldsymbol{D}$ 的大小为 n，$\boldsymbol{D}$ 中全部词语的语义向量构成语义向量集合 $\boldsymbol{S}$，$\boldsymbol{S}=\{\boldsymbol{x}_i:\boldsymbol{x}_i=(x_{i1},\cdots,x_{ij},\cdots,x_{id})$；$\boldsymbol{x}_i\in\boldsymbol{R}^d$；$x_{ij}\in\boldsymbol{R}$；$i=1,\cdots,n$；$j=1,\cdots,d\}$，$\boldsymbol{S}$ 为 $\boldsymbol{D}$ 对应的语义向量空间。将语义向量空间 $\boldsymbol{S}$ 看作总体，所有语义向量 $\boldsymbol{x}$ 的第 j 维 x_j 构成的向量记为 $\boldsymbol{y}_j$，称为维度向量，$\boldsymbol{y}_j=(x_{1j},\cdots,x_{ij},\cdots,x_{nj})$。将 $\boldsymbol{y}_j$ 看作统计样本，在本文第 4 节中的实验数据 DataSet1 上得到的 10 维语义向量的各个维度的统计直方图如图 1 所示，在实验数据 DataSet2 上得到的语义向量的统计分布与此类似，不再赘述。

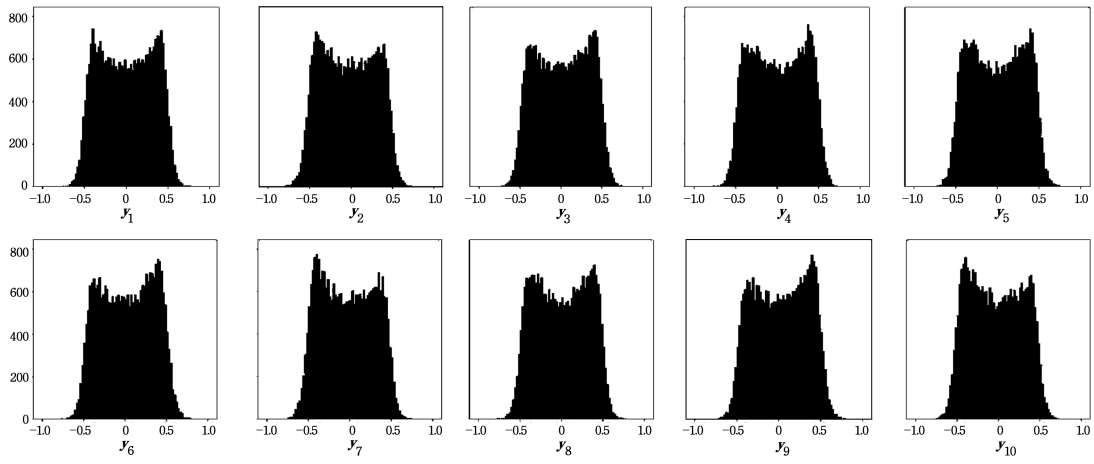

图 1 10 维语义向量的 10 个维度的直方图

Fig. 1 Histogram of 10 dimensions of 10-dimensional semantic vector

图 1 中，$\boldsymbol{y}_1$ 至 $\boldsymbol{y}_{10}$ 为语义向量的 10 个维度向量，横轴为维度值，纵轴为维度值的频率。从图 1 中可得出 3 点结论：1）不同的维度具有近似的统计分布。2）维度向量 $\boldsymbol{y}_j$ 没有明显的统计分布规律，其分布不属于常见的概率分布，很难拟合出准确的概率密度函数，因此采用矩估计法通过样本估计统计特征值；如果读者所用的语义向量的维度向量的概率密度函数可以较容易地通过拟合得到，那么统计特征值可由概率密度函数计算。3）维度向量 $\boldsymbol{y}_j$ 的元素的取值范围在[−0.8587，0.8663]之间，均值范围在[−0.0299，0.0318]之间，标准差最大为0.3168；由第 4 节的实验数据可知，样本为 28565 个，这些样本数据集中在[−0.8587，0.8677]范围内，因此会有大量的维度值近似相等，而距离由维度值的差异累积产生，所以距离会相等或近似相等，从而出现“度量集中”现象。

当语义向量的维度从 10 维增加至 400 维时，不同维度向量的直方图形状变化情况如图 2 所示。

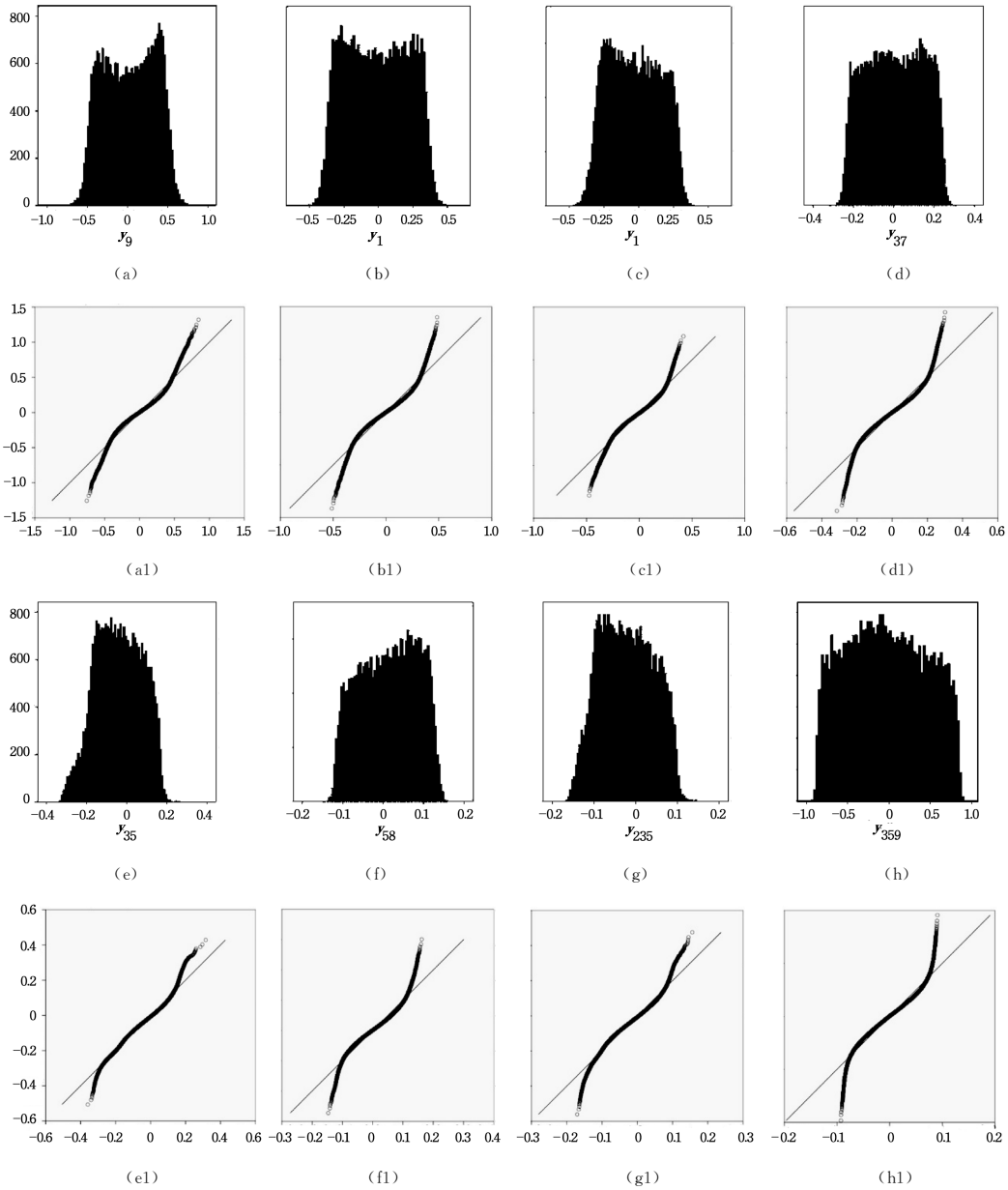

图 2　不同维度向量的直方图和 Q-Q 统计图

Fig. 2　Histograms and Q-Q charts of different dimension vectors

图 2 给出了从不同维度的语义向量中随机选取一个维度作为结果展示效果。随着维度的增加，直方图的形状类似于正态分布，因此除了直方图外，还给出了用于检验是否为正态分布的 Q-Q 统计图，以进一步说明统计分布规律。

由图 2 可知，不同维度向量的分布不尽相同。从直方图和 Q-Q 统计图中得出，维度向量没有明显的统计分布规律，在显著性取值为 0.05 下的 k-s 检验和卡方检验都拒绝了正态分布假设；由于随机变量取值范围中有负数，因此不是卡方分布、指数分布、对数正态分布、伽玛分布，统计特征值由样本估计得到。

随着维度增加，语义向量维度值的取值范围逐渐缩小，如图 3 所示。最大标准差也逐渐缩小，如图 4 所示。均值范围

具有略微缩小的趋势，但变化不明显，如图 5 所示。

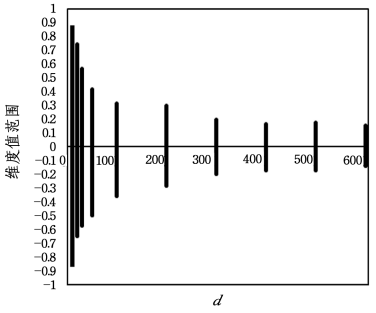

图 3 维度值范围变化情况

Fig. 3 Variation range of dimension values

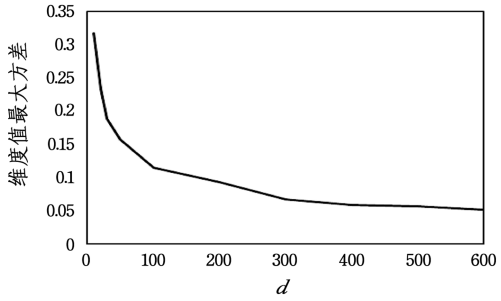

图 4 维度值最大方差变化曲线图

Fig. 4 Variation curve of maximum variance of dimension values

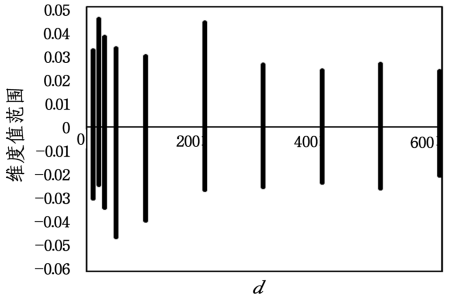

图 5 均值范围变化情况

Fig. 5 Variation range of average dimension values

因此当维度增加时，“度量集中”问题表现得更加明显。

3 差异性度量方法

杰卡德系数广泛应用于数据的相似性和差异性度量中[5,19-21]，除自然语言处理领域的应用外[5]，它在其他领域中也取得了较好的应用效果[20-21]。本文以非对称二值特征杰卡德系数为基础，阐述基于该系数的差异性度量方法，然后分析该度量方法在度量高维语义向量时的局限性，进而阐述基于非对称多值特征杰卡德系数的差异性度量方法。

3.1 语义向量离散化

杰卡德系数是一种度量离散数据相似度的方法，因此本文首先将语义向量进行离散化处理。离散化过程的本质是一种非线性变换，理论上，这种变换有利于对非线性数据关系进行描述，而语义关系属于非线性关系，因此这种离散化处理能简化应用语义向量的模型，进而改善应用效果。

一般的语义向量 $\boldsymbol{x}\in\boldsymbol{R}^d$，第 j 维 $x_j\in\boldsymbol{R}$，而利用语言模型训练深度神经网络得到的语义向量 $\boldsymbol{x}\in[-1,1]^d$，$x_j\in[-1,1]$。根据统计分析结果（见第 2 节）可知，随着 $\boldsymbol{x}$ 的维度 d 的增加，x_j 的区间缩小。设 x_j 的区间为 $[a,b]$，$-1\leqslant a<0$，$0<b\leqslant 1$，本文采用区间 $[a,b]$ 内等距的离散化方法，将区间 $[a,b]$ 均匀划分为 $c(c\geqslant 2)$ 份，c 取整数。得到的离散化语义向量 $\boldsymbol{x}^{\#}=(x_1^{\#},\cdots,x_j^{\#},\cdots,x_d^{\#})$，$\boldsymbol{x}^{\#}\in\boldsymbol{Z}^d$，$x_j^{\#}\in\boldsymbol{Z}$，$\boldsymbol{Z}=\{0,\cdots,c-1\}$。具体的离散化过程为：给定 x_j，当

$$2z/c-1\leqslant x_j<2(z+1)/c-1 \tag{1}$$

成立时，$x_j^{\#}=z$，其中，$z\in\boldsymbol{Z}$。当 $c=2$ 时，语义向量 $\boldsymbol{x}$ 被离散化为 0-1 二值语义向量，将离散化的 $\boldsymbol{S}$ 记为 $\boldsymbol{S}^{\#}$。

3.2 基于非对称二值特征杰卡德系数的度量方法

在 $\boldsymbol{S}$ 上任取两个语义向量，记为 $\boldsymbol{u}$ 和 $\boldsymbol{v}$，分别为词表 $\boldsymbol{D}$ 中的词语 w_u 和 w_v 的语义向量，离散化后的语义向量记为 $\boldsymbol{u}^{\#}$ 和 $\boldsymbol{v}^{\#}$，且 $\boldsymbol{u}^{\#},\boldsymbol{v}^{\#}\in\boldsymbol{S}^{\#}$。$\boldsymbol{u}$ 和 $\boldsymbol{v}$ 的差异性度量结果用 $\boldsymbol{u}$ 和 $\boldsymbol{v}$ 的距离表示，记为 $d(\boldsymbol{u},\boldsymbol{v})$，本文令 $d(\boldsymbol{u},\boldsymbol{v})=d(\boldsymbol{u}^{\#},\boldsymbol{v}^{\#})$。

语义向量 $\boldsymbol{u}^{\#}$ 和 $\boldsymbol{v}^{\#}$ 的非对称二值特征杰卡德系数定义如式(2)所示。

$$J(\boldsymbol{u}^{\#},\boldsymbol{v}^{\#})=\frac{M_{11}}{M_{01}+M_{10}+M_{11}} \tag{2}$$

其中，$J(\boldsymbol{u}^{\#},\boldsymbol{v}^{\#})$为二值特征杰卡德系数，$\boldsymbol{u}^{\#},\boldsymbol{v}^{\#}$ 为 0-1 二值语义向量，M_{11} 表示 $\boldsymbol{u}^{\#}$ 和 $\boldsymbol{v}^{\#}$ 的维度值同时为 1 的情况的数目；M_{01} 表示 $\boldsymbol{u}^{\#}$ 的维度值为 0 且 $\boldsymbol{v}^{\#}$ 的维度值为 1 的情况的数目；M_{10} 表示 $\boldsymbol{u}^{\#}$ 的维度值为 1 且 $\boldsymbol{v}^{\#}$ 的维度值为 0 的情况的数目。因此，M_{11} 可由式(3)表示。

$$M_{11}=\sum_{j=1}^{d}u_jv_j=\boldsymbol{u}\odot\boldsymbol{v}=\sum_{j=1}^{d}u_j\wedge v_j \tag{3}$$

其中，$\odot$ 表示向量的内积运算，$\wedge$ 表示逻辑与运算。

式(2)中 $M_{01}+M_{10}+M_{11}$ 可由式(4)表示：

$$M_{01}+M_{10}+M_{11}=\sum_{j=1}^{d}u_j\vee v_j \tag{4}$$

其中，$\vee$ 表示逻辑或运算。

式(2)－式(4)满足以下 5 个约束条件：

1) $0\leqslant M_{11}\leqslant d$；

2) $0\leqslant M_{01}+M_{10}+M_{11}\leqslant d$；

3) $0\leqslant M_{01}\leqslant d$；

4) $0\leqslant M_{10}\leqslant d$；

5) $M_{01}+M_{10}+M_{11}+M_{00}=d$。

由式(2)－式(4)和以上 5 个约束条件可证，$J(\boldsymbol{u}^{\#},\boldsymbol{v}^{\#})\in[0,1]$。当式(2)中 M_{11}，M_{01} 和 M_{10} 同时为 0，即 $\boldsymbol{u}^{\#}$ 和 $\boldsymbol{v}^{\#}$ 为零向量时，为保证该方法具有良定义，规定 $J(\boldsymbol{u}^{\#},\boldsymbol{v}^{\#})=0$。

基于非对称二值特征杰卡德系数的差异性度量方法如式(5)所示：

$$d(\boldsymbol{u},\boldsymbol{v})=d(\boldsymbol{u}^{\#},\boldsymbol{v}^{\#})=F-J(\boldsymbol{u}^{\#},\boldsymbol{v}^{\#}) \tag{5}$$

该度量方法记为 J_Dist。

度量方法一般应满足度量空间上的正定性、对称性和三角不等式 3 个条件。对于度量方法 J_Dist，由于式(5)中的 $J(\boldsymbol{u}^{\#},\boldsymbol{v}^{\#})\in[0,1]$，因此取 $F=1$ 时，对于 $\forall\boldsymbol{u}$ 和 $\forall\boldsymbol{v}$，都满足 $0\leqslant d(\boldsymbol{u},\boldsymbol{v})\leqslant 1$，即满足正定性。式(5)中，当且仅当 $\boldsymbol{u}^{\#}=\boldsymbol{v}^{\#}\neq\boldsymbol{0}$ 成立时，$d(\boldsymbol{u},\boldsymbol{v})=0$；当 $\boldsymbol{u}^{\#}=\boldsymbol{v}^{\#}=\boldsymbol{0}$ 成立时，$d(\boldsymbol{u},\boldsymbol{v})=1$，即式(2)中 M_{11}，M_{01} 和 M_{10} 同时为 0。易知，J_Dist 满足对称性，即 $\forall\boldsymbol{u}$ 和 $\forall\boldsymbol{v}$，$d(\boldsymbol{u},\boldsymbol{v})=d(\boldsymbol{v},\boldsymbol{u})$。

由式(2)－式(4)联合推导出式(6)：

$$J(\boldsymbol{u}^{\#},\boldsymbol{v}^{\#})=\frac{\sum_{j=1}^{d}u_j\wedge v_j}{\sum_{j=1}^{d}u_j\vee v_j} \tag{6}$$

由文献[22-23]可知，式(6)为谷本系数(Tanimoto Coefficient)，式(5)为谷本距离(Tanimoto Distance)。由文献[18，22-23]可知，当 $\boldsymbol{u}^{\#}$ 和 $\boldsymbol{v}^{\#}$ 为二值特征向量时，度量方法 J_Dist 满足三角不等式，即 $\forall \boldsymbol{u},\forall \boldsymbol{v},\forall \boldsymbol{q},\boldsymbol{u},\boldsymbol{v},\boldsymbol{q}\in \boldsymbol{S}$，$d(\boldsymbol{u},\boldsymbol{q})\leqslant d(\boldsymbol{u},\boldsymbol{v})+d(\boldsymbol{v},\boldsymbol{q})$。

从式(2)－式(5)可以得出，J_Dist 方法表达维度值同时为 1 的数目与总维度 d 中去除维度值都为 0 的情况的数目的比值，因此该方法计算的是非对称的二值向量的相似度或差异度，0 值和 1 值不是同等重要的，由重要的值决定差异度。一般地，将出现概率较小的维度值编码为 1，将另一种维度值编码为 0，即认为小概率事件对差异性的影响更大。因此该方法将语义向量的各个维度视为非同等重要，值为 1 的维度贡献了更多的差异度。

3.3　基于非对称多值特征杰卡德差异性的度量方法

J_Dist 方法要求语义向量为二值向量，且值为 1 的维度贡献更多的差异度，但该方法存在两个问题：1)式(1)中 $c>2$ 时离散语义向量为多值向量，J_Dist 方法无法处理；2)若 $c>2$，如何判定哪些维度贡献更多的差异度。基于这两个问题，提出一种基于非对称多值特征杰卡德系数的度量方法，该方法的直觉意义为：若维度值为 $z(z\in \boldsymbol{Z})$ 的频率很高，甚至接近于 n，则该维度贡献的差异度很小，弱化该类维度对差异度的贡献量可提高度量方法的差异度，因此该方法定义关于维度值频率的重要性函数，由重要性函数值决定其贡献的差异度，具体步骤阐述如下。

步骤 1　统计 $\boldsymbol{S}^{\#}$ 中语义向量 $\boldsymbol{x}^{\#}$ 的第 j 维 $x_j^{\#}$ 的值为 z 的频率，记为 $freq_{j,z}$，则 $\boldsymbol{S}^{\#}$ 中所有语义向量的维度值为 z $(z\in \boldsymbol{Z})$ 的频率记为矩阵 $\boldsymbol{Freq}$，$\boldsymbol{Freq}=[freq_{j,z}]_{d*c}$。

步骤 2　定义由 $\boldsymbol{Freq}$ 确定的重要性函数 $f(\boldsymbol{Freq})=[f(freq_{j,z})]_{d*c}$，在实际应用中建议该函数选择单调函数，例如，$f(freq_{j,z})=1/freq_{j,z}$。

步骤 3　根据函数 $f(\boldsymbol{Freq})$ 的值确定贡献语义向量差异度的维度值集合。设第 j 维的阈值记为 $\boldsymbol{T}_j^f$，则所有维度的阈值构成向量 $\boldsymbol{T}^f=(T_1^f,\cdots,T_j^f,\cdots,T_d^f)$。满足 $f(freq_{j,z})>T_j^f$ 的 $x_j^{\#}$ 的值 z 构成的集合为 Z_j^f，则所有满足 $f(freq_{j,z})>T_j^f$ 的 $x_j^{\#}$ 的值 z 构成向量 $\boldsymbol{Z}^f=(Z_1^f,\cdots,Z_j^f,\cdots,Z_d^f)$，$Z_j^f\subset \boldsymbol{Z}$；不满足 $f(freq_{j,z})>T_j^f$ 的 $x_j^{\#}$ 的值 z 构成的集合记为 $\overline{Z_j^f}$，则所有维度的维度值不满足 $f(freq_{j,z})>T_j^f$ 的值 z 构成向量 $\overline{\boldsymbol{Z}^f}=(\overline{Z_1^f},\cdots,\overline{Z_j^f},\cdots,\overline{Z_d^f})=(\boldsymbol{Z}-Z_1^f,\cdots,\boldsymbol{Z}-Z_j^f,\cdots,\boldsymbol{Z}-Z_d^f)$，$\boldsymbol{Z}-Z_j^f$ 表示集合 $\boldsymbol{Z}$ 与集合 Z_j^f 的差集，$\overline{Z_j^f}\subset \boldsymbol{Z}$，$Z_j^f\cap\overline{Z_j^f}=\emptyset$。

步骤 4　将 $\boldsymbol{u}^{\#}$，$\boldsymbol{v}^{\#}$ 中满足 $u_j^{\#}\in Z_j^f$ 且 $v_j^{\#}\in Z_j^f$ 的维度作为计算多值特征杰卡德系数的主要依据，忽略 $u_j^{\#}\in\overline{Z_j^f}$ 且 $v_j^{\#}\in\overline{Z_j^f}$ 的维度。在式(2)的基础上，给出多值特征特征杰卡德系数，记为 $JM(\boldsymbol{u}^{\#},\boldsymbol{v}^{\#})$，如式(7)所示：

$$JM(\boldsymbol{u}^{\#},\boldsymbol{v}^{\#})=\frac{M_{\boldsymbol{Z}^f,\boldsymbol{Z}^f}}{M_{\overline{\boldsymbol{Z}^f},\boldsymbol{Z}^f}+M_{\boldsymbol{Z}^f,\overline{\boldsymbol{Z}^f}}+M_{\boldsymbol{Z}^f,\boldsymbol{Z}^f}} \tag{7}$$

步骤 5　由式(7)给出语义向量 $\boldsymbol{u}$ 和 $\boldsymbol{v}$ 的差异性度量方法，如式(8)所示：

$$d(\boldsymbol{u},\boldsymbol{v})=d(\boldsymbol{u}^{\#},\boldsymbol{v}^{\#})=F-JM(\boldsymbol{u}^{\#},\boldsymbol{v}^{\#}) \tag{8}$$

至此，该方法阐述完毕，该方法记为 JM_Dist。该方法中式(7)各符号的计算方法和式(8)中 F 的取值阐述如下。

式(7)中，$M_{\boldsymbol{Z}^f,\boldsymbol{Z}^f}$ 的计算方法如式(9)所示：

$$M_{\boldsymbol{Z}^f,\boldsymbol{Z}^f}=\sum_{j=1}^{d}\frac{1}{|u_j^{\#}-v_j^{\#}|+1}{}_{u_j^{\#}\in Z_j^f,v_j^{\#}\in Z_j^f} \tag{9}$$

式(9)表示条件 $u_j^{\#}\in Z_j^f$ 和 $v_j^{\#}\in Z_j^f$ 成立时计算 $1/(|u_j^{\#}-v_j^{\#}|+1)$；否则不计算 $1/(|u_j^{\#}-v_j^{\#}|+1)$，计算后累加求和。

$M_{\overline{\boldsymbol{Z}^f},\boldsymbol{Z}^f}$ 的计算方法如式(10)所示：

$$M_{\overline{\boldsymbol{Z}^f},\boldsymbol{Z}^f}=\sum_{j=1}^{d}\frac{1}{|u_j^{\#}-v_j^{\#}|+1}{}_{u_j^{\#}\in \overline{Z_j^f},v_j^{\#}\in Z_j^f} \tag{10}$$

式(10)表示条件 $u_j^{\#}\in\overline{Z_j^f}$ 和 $v_j^{\#}\in Z_j^f$ 成立时计算 $1/(|u_j^{\#}-v_j^{\#}|+1)$；否则不计算 $1/(|u_j^{\#}-v_j^{\#}|+1)$，计算后累加求和。

$M_{\boldsymbol{Z}^f,\overline{\boldsymbol{Z}^f}}$ 的计算方法如式(11)所示：

$$M_{\boldsymbol{Z}^f,\overline{\boldsymbol{Z}^f}}=\sum_{j=1}^{d}\frac{1}{|u_j^{\#}-v_j^{\#}|+1}{}_{u_j^{\#}\in Z_j^f,v_j^{\#}\in \overline{Z_j^f}} \tag{11}$$

式(11)表示条件 $u_j^{\#}\in Z_j^f$ 和 $v_j^{\#}\in\overline{Z_j^f}$ 成立时计算 $1/(|u_j^{\#}-v_j^{\#}|+1)$；否则不计算 $1/(|u_j^{\#}-v_j^{\#}|+1)$，计算后累加求和。

分析式(9)－式(11)，$|u_j^{\#}-v_j^{\#}|$ 表达语义向量的分量的差异，而杰卡德系数本身表达相似性，因此取 $1/|u_j^{\#}-v_j^{\#}|$ 作为计算相似度的依据，但为避免语义向量对应的分量完全相同时 $|u_j^{\#}-v_j^{\#}|=0$ 使得 $1/|u_j^{\#}-v_j^{\#}|$ 出现不适定的情况，令 $1/(|u_j^{\#}-v_j^{\#}|+1)$ 为计算相似度的依据。当 $|u_j^{\#}-v_j^{\#}|=0$ 时，式(9)中 $M_{\boldsymbol{Z}^f,\boldsymbol{Z}^f}=1$。而 $M_{\overline{\boldsymbol{Z}^f},\boldsymbol{Z}^f}$ 和 $M_{\boldsymbol{Z}^f,\overline{\boldsymbol{Z}^f}}$ 中 $u_j^{\#}$ 和 $v_j^{\#}$ 分别属于互不相交的两个集合 $\overline{Z_j^f}$ 和 Z_j^f，因此式(10)和式(11)中 $|u_j^{\#}-v_j^{\#}|\neq 0$ 必定成立。

当 $\boldsymbol{u}^{\#}$ 的所有维度值 $u_j^{\#}\in\overline{Z_j^f}$ 且 $\boldsymbol{v}^{\#}$ 的所有对应的维度值 $v_j^{\#}\in\overline{Z_j^f}$ 成立时，$JM(\boldsymbol{u}^{\#},\boldsymbol{v}^{\#})$ 无法计算，为满足方法的完备性，规定 $JM(\boldsymbol{u}^{\#},\boldsymbol{v}^{\#})=0$。此情况表示两个语义向量的所有对应的维度都不具有体现语义向量差异的能力，与式(2)中 M_{11}，M_{01} 和 M_{10} 同时为 0 时的情况类似，认为语义向量为零向量。从语义角度分析，这两个语义向量表示的词语不具有相似的语义，即语义相似系数为零。

由以上分析得出，式(7)、式(9)－式(11)满足以下 4 个约束条件：

1)$0\leqslant M_{\boldsymbol{Z}^f,\boldsymbol{Z}^f}\leqslant d$；

2)$0\leqslant M_{\overline{\boldsymbol{Z}^f},\boldsymbol{Z}^f}\leqslant d/2$；

3)$0\leqslant M_{\boldsymbol{Z}^f,\overline{\boldsymbol{Z}^f}}\leqslant d/2$；

4)$0\leqslant M_{\overline{\boldsymbol{Z}^f},\boldsymbol{Z}^f}+M_{\boldsymbol{Z}^f,\overline{\boldsymbol{Z}^f}}+M_{\boldsymbol{Z}^f,\boldsymbol{Z}^f}\leqslant d$。

由式(7)、式(9)－式(11)和以上 4 个约束条件可证，$0\leqslant JM(\boldsymbol{u}^{\#},\boldsymbol{v}^{\#})\leqslant 1$。

由于式(8)中 $0\leqslant JM(\boldsymbol{u}^{\#},\boldsymbol{v}^{\#})\leqslant 1$，为满足度量方法的正定性，式(8)中取 $F=1$，对于 $\forall \boldsymbol{u},\forall \boldsymbol{v}$，都满足 $0\leqslant d(\boldsymbol{u},\boldsymbol{v})\leqslant 1$，

即该方法满足正定性。易证,JM_Dist 方法满足对称性,即对于 $\forall \boldsymbol{u},\forall \boldsymbol{v}, d(\boldsymbol{u},\boldsymbol{v})=d(\boldsymbol{v},\boldsymbol{u})$。通过实验可知 JM_Dist 不满足三角不等式,理论证明和分析过程如下。

已知:$\boldsymbol{u},\boldsymbol{v}$ 和 $\boldsymbol{q}$ 为 $\boldsymbol{S}$ 中任意 3 个语义向量,$d(\boldsymbol{u},\boldsymbol{v})$如式(8)所示,$JM(\boldsymbol{u}^{\#},\boldsymbol{v}^{\#})$如式(7)、式(9)—式(11)所示。

求证:

$$d(\boldsymbol{u},\boldsymbol{v})+d(\boldsymbol{u},\boldsymbol{q})>d(\boldsymbol{v},\boldsymbol{q}) \tag{12}$$

$$d(\boldsymbol{u},\boldsymbol{q})+d(\boldsymbol{v},\boldsymbol{q})>d(\boldsymbol{u},\boldsymbol{v}) \tag{13}$$

$$d(\boldsymbol{v},\boldsymbol{q})+d(\boldsymbol{u},\boldsymbol{v})>d(\boldsymbol{u},\boldsymbol{q}) \tag{14}$$

以上 3 个不等式不成立。

证明:只证明不等式(12)不成立,式(13)和式(14)的证明方法与式(12)相同。

$$d(\boldsymbol{u},\boldsymbol{v})+d(\boldsymbol{u},\boldsymbol{q})=1-JM(\boldsymbol{u}^{\#},\boldsymbol{v}^{\#})+1-JM(\boldsymbol{u}^{\#},\boldsymbol{q}^{\#}) \tag{15}$$

$$d(\boldsymbol{v},\boldsymbol{q})=1-JM(\boldsymbol{v}^{\#},\boldsymbol{q}^{\#}) \tag{16}$$

将式(15)和式(16)代入式(12),整理得:

$$JM(\boldsymbol{u}^{\#},\boldsymbol{v}^{\#})+JM(\boldsymbol{u}^{\#},\boldsymbol{q}^{\#})-JM(\boldsymbol{v}^{\#},\boldsymbol{q}^{\#})<1 \tag{17}$$

按式(7)展开式(17)中的 $JM(\boldsymbol{u}^{\#},\boldsymbol{v}^{\#})$,$JM(\boldsymbol{u}^{\#},\boldsymbol{q}^{\#})$和 $JM(\boldsymbol{v}^{\#},\boldsymbol{q}^{\#})$,得到式(18):

$$\frac{M^{\boldsymbol{uv}}_{\boldsymbol{Z}^f,\boldsymbol{Z}^f}}{M^{\boldsymbol{uv}}_{\boldsymbol{Z}^f,\boldsymbol{Z}^f}+M^{\boldsymbol{uv}}_{\boldsymbol{Z}^f,\overline{\boldsymbol{Z}^f}}+M^{\boldsymbol{uv}}_{\overline{\boldsymbol{Z}^f},\boldsymbol{Z}^f}}+\frac{M^{\boldsymbol{uq}}_{\boldsymbol{Z}^f,\boldsymbol{Z}^f}}{M^{\boldsymbol{uq}}_{\overline{\boldsymbol{Z}^f},\boldsymbol{Z}^f}+M^{\boldsymbol{uq}}_{\boldsymbol{Z}^f,\overline{\boldsymbol{Z}^f}}+M^{\boldsymbol{uq}}_{\boldsymbol{Z}^f,\boldsymbol{Z}^f}}-\frac{M^{\boldsymbol{vq}}_{\boldsymbol{Z}^f,\boldsymbol{Z}^f}}{M^{\boldsymbol{vq}}_{\overline{\boldsymbol{Z}^f},\boldsymbol{Z}^f}+M^{\boldsymbol{vq}}_{\boldsymbol{Z}^f,\overline{\boldsymbol{Z}^f}}+M^{\boldsymbol{vq}}_{\boldsymbol{Z}^f,\boldsymbol{Z}^f}}<1 \tag{18}$$

分析式(9)—式(11)中的条件 $u_j^{\#}\in Z_j^f$,$v_j^{\#}\in Z_j^f$,$u_j^{\#}\in\overline{Z_j^f}$ 和 $v_j^{\#}\in\overline{Z_j^f}$可知:

$\because Z_j^f\cap\overline{Z_j^f}=\varnothing$

$\therefore$条件 $u_j^{\#}\in Z_j^f$ 且 $v_j^{\#}\in Z_j^f$ 与条件 $u_j^{\#}\in\overline{Z_j^f}$且 $v_j^{\#}\in Z_j^f$ 不可能同时成立。

$\therefore j$ 取某一值时,若式(9)中 $1/(|u_j^{\#}-v_j^{\#}|+1)$的条件成立,则式(10)的条件一定不成立;同理,式(11)中的条件也一定不成立,不成立时 $1/(|u_j^{\#}-v_j^{\#}|+1)$不参与差异性累积计算。同理可推断出,式(9)—式(11)中若有一个条件成立,则另外两个公式的条件不成立。为化简式(18),引入特征函数 τ,结合上述分析,对于 $d(\boldsymbol{u},\boldsymbol{v})$,式(9)中条件成立时 $\tau(u_j^{\#},v_j^{\#})=1$,式(10)和式(11)中 $\tau(u_j^{\#},v_j^{\#})=0$。$d(\boldsymbol{u},\boldsymbol{q})$和 $d(\boldsymbol{v},\boldsymbol{q})$的情况同上,略。

$\therefore$式(9)可表示为:

$$M_{\boldsymbol{Z}^f,\boldsymbol{Z}^f}=\sum_{j=1}^{d}\frac{1}{|u_j^{\#}-v_j^{\#}|+1}\tau(u_j^{\#},v_j^{\#})$$

同理可得 $M_{\overline{\boldsymbol{Z}^f},\boldsymbol{Z}^f}$和 $M_{\boldsymbol{Z}^f,\overline{\boldsymbol{Z}^f}}$,具体公式略。

$\therefore$式(18)的第一项的分母 $M^{\boldsymbol{uv}}_{\overline{\boldsymbol{Z}^f},\boldsymbol{Z}^f}+M^{\boldsymbol{uv}}_{\boldsymbol{Z}^f,\overline{\boldsymbol{Z}^f}}+M^{\boldsymbol{uv}}_{\boldsymbol{Z}^f,\boldsymbol{Z}^f}=\sum_{j=1}^{d}\frac{1}{|u_j^{\#}-v_j^{\#}|+1}$。

$\therefore$式(18)可转化为式(19):

$$\frac{\sum_{j=1}^{d}\frac{1}{|u_j^{\#}-v_j^{\#}|+1}\tau(u_j^{\#},v_j^{\#})}{\sum_{j=1}^{d}\frac{1}{|u_j^{\#}-v_j^{\#}|+1}}+\frac{\sum_{j=1}^{d}\frac{1}{|u_j^{\#}-q_j^{\#}|+1}\tau(u_j^{\#},q_j^{\#})}{\sum_{j=1}^{d}\frac{1}{|u_j^{\#}-q_j^{\#}|+1}}-\frac{\sum_{j=1}^{d}\frac{1}{|v_j^{\#}-q_j^{\#}|+1}\tau(v_j^{\#},q_j^{\#})}{\sum_{j=1}^{d}\frac{1}{|v_j^{\#}-q_j^{\#}|+1}}<1 \tag{19}$$

其中,分母为确定的表达形式,分子中包含特征函数 τ,τ 为不确定的表达形式。特征函数 $\tau(u_j^{\#},v_j^{\#})$,$\tau(u_j^{\#},q_j^{\#})$和 $\tau(v_j^{\#},q_j^{\#})$之间不存在约束或比较关系,因此找不到确定的等式或不等式进行代换,无法进行严格的不等式推导。1)由前文可知,差异性累积算式 $1/(|u_j^{\#}-v_j^{\#}|+1)$由分量的频率值的函数 $f(\boldsymbol{Freq})$确定,当 $f(\boldsymbol{Freq})$取反比例函数时,频率值越高,贡献的差异度越小;2)频率值越高,$|u_j^{\#}-v_j^{\#}|$不一定越小,举例说明如下。

不失一般性,以参数 $c=10$,$d=3$ 为例,语义向量 $\boldsymbol{u}^{\#}=(0,8,7)$,$\boldsymbol{v}^{\#}=(1,6,2)$,$\boldsymbol{q}^{\#}=(5,8,6)$,$\boldsymbol{u}^{\#},\boldsymbol{v}^{\#},\boldsymbol{q}^{\#}\in\boldsymbol{S}^{\#}$。根据函数 $f(\boldsymbol{Freq})$和阈值向量 $\boldsymbol{T}^f$ 计算可得 $Z_1^f=\{2,3,5,7,8\}$,$\overline{Z_1^f}=\{0,1,4,6,9\}$;$Z_2^f=\{0,2,6,7,8,9\}$,$\overline{Z_2^f}=\{1,3,4,5\}$;$Z_3^f=\{1,3,4,6,7\}$,$\overline{Z_3^f}=\{0,2,5,8,9\}$。此时,$u_2^{\#}\in z_2^f$ 且 $v_2^{\#}\in z_2^f$,因此 $|u_2^{\#}-v_2^{\#}|=|8-6|=2$,分量 $u_2^{\#}$ 和 $v_2^{\#}$ 属于频率较低的维度;$u_3^{\#}\in z_3^f$ 且 $v_3^{\#}\in z_3^f$,因此 $|u_3^{\#}-v_3^{\#}|=|7-2|=5$,分量 $v_3^{\#}$ 属于频率较高的维度,反而得到了较大的差异度。综合 1)、2)两点得出:差异性的大小与频率的高低无相关关系,因此式(19)难以继续进行严格的不等式推导。将上述数据代入式(19)计算可得不成立,即验证了不等式 $d(\boldsymbol{u},\boldsymbol{v})+d(\boldsymbol{u},\boldsymbol{q})<d(\boldsymbol{v},\boldsymbol{q})$不成立,同理式(13)和式(14)也不成立,因此 J_Dist不满足三角不等式。

JM_Dist 方法解决了 J_Dist 方法存在的两个问题:1)JM_Dist方法中 $\boldsymbol{u}^{\#}$ 和 $\boldsymbol{v}^{\#}$ 为多值向量,差异度由 $|u_j^{\#}-v_j^{\#}|$ 计算,较 0-1 二值向量有更细粒度的差异性比较;2)根据步骤 3 和步骤 4,将语义向量 $\boldsymbol{u}^{\#}$ 和 $\boldsymbol{v}^{\#}$ 中满足 $f(freq_{j,z})>T_j^f$ 的维度作为多值特征杰卡德距离的主要依据,忽略了无法贡献差异性的维度,降低了语义向量的维度。如果式(1)中 c 取值适当且步骤 2 中的函数 $f(freq_{j,z})$设计合理,则该方法可有效缓解“度量集中”问题。

4 实验结果及分析

实验数据为语义向量集合 $\boldsymbol{S}$,它通过语言模型训练深度神经网络而得到,因此首先阐述训练神经网络模型的参数。为评价度量方法对“度量集中”问题的缓解程度,本文给出两个指标:多样性和语义性。实验中给出了不同度量方法的多样性和语义性,并进行了分析。

4.1 实验数据

为体现训练数据具有“小样本”问题且得到的语义向量的度量方法存在“度量集中”现象,本文未选取大规模的语料,而是选取渔业领域的《水产辞典》[24]中的全部文本和搜狗全网新闻数据 SogouCA[25]作为训练语料。前者记为 DataSet1,该语料具有较强的语义含义,适合检验度量方法的语义性指标;后者记为 DataSet2,该数据为公开数据集,可验证本文提出的度量方法更具普遍意义。数据集 DataSet1 和 DataSet2 中包含的词语数量等信息如表 1 所列。其中,词语数量为对数据进行分词和去除停用词后的词语数量。

表 1　实验数据

Table 1　Experimental data

数据集	词语数量/MB	训练工具	训练模型	词表大小/kB	维度 d	模型参数
DataSet1	0.20	Word2Vec	Skip-gram	27.90	10,20,30,50,100,200,300,400	$\eta=0.0015$；$sample=1.0$；$window=10$；优化方法为 Hierarchical Softmax
DataSet2	250	Word2Vec	CBOW	752	10,20,30,50,100,200,300,400	$\eta=0.015$；$sample=0.1$；$window=5$；优化方法为 Negative Sampling

两个数据集的训练工具都选择 Word2Vec[6-7]。根据文献[16]，当数据规模较小时采用 Skip-gram 模型进行训练；当数据规模较大时采用 CBOW 模型进行训练。因此，DataSet1 选择 Skip-gram 模型，DataSet2 选择 CBOW 模型。模型参数中，η 表示学习率，$sample$ 表示采样频率，$window$ 表示窗口大小。为提高训练效率，在神经网络的输出层分别采用 Hierarchical Softmax 和 Negative Sampling 方法进行优化。为验证不同维度下“度量集中”问题的表现程度和本文的度量方法的效果，在两个数据集上分别训练了 $d=10,20,30,50,100,200,300$ 和 400 维度的语义向量，词表 $\boldsymbol{D}$ 的词语个数为 n，因此 $\boldsymbol{S}$ 为 $n*d$ 的矩阵，将其作为差异性度量方法的实验数据。

4.2　评价指标

若通过某种度量方法得到的语义向量间的距离具有较大的差异，则认为距离具有多样性，缓解了“度量集中”问题，因此多样性指标是本文方法的评判依据之一，具体评价算法在 4.2.1节阐述。由于本文的方法应用于具有语义含义的语义向量中，因此在提高多样性的同时是否能够准确地表达语义差异是重要的评判依据之二，具体评价方法在 4.2.2 节阐述。

4.2.1　多样性指标

多样性由语义向量距离的均值和标准差两个统计量表达，具体算法如下。

首先，计算 $\boldsymbol{S}$ 中任意两个语义向量 $\boldsymbol{x}_{i_1}$ 和 $\boldsymbol{x}_{i_2}$ 间的距离 $d(\boldsymbol{x}_{i_1},\boldsymbol{x}_{i_2})$，得到的距离矩阵 $\boldsymbol{dist}(\boldsymbol{S})$ 如式(20)所示。

$$\boldsymbol{dist}(\boldsymbol{S})=[d(\boldsymbol{x}_{i_1},\boldsymbol{x}_{i_2})]_{n*n}=\begin{bmatrix} d(x_1,x_1) & \cdots & d(x_1,x_n) \\ \vdots & \cdots & \vdots \\ d(x_n,x_1) & \cdots & d(x_n,x_n) \end{bmatrix} \tag{20}$$

其次，计算均值和标准差。由于度量方法具有对称性，因此式(20)中的矩阵是主对角线元素为 0 的对称矩阵，只需计算主对角线以上的 $n(n-1)/2$ 个元素的均值 μ_{dist} 和标准差 σ_{dist}，如式(21)和式(22)所示：

$$\mu_{dist}=\frac{2}{n(n-1)}\sum_{i_1<i_2} d(\boldsymbol{x}_{i_1},\boldsymbol{x}_{i_2}) \tag{21}$$

$$\sigma_{dist}=\sqrt{\sum_{i_1<i_2}(d(\boldsymbol{x}_{i_1},\boldsymbol{x}_{i_2})-\mu_{dist})^2} \tag{22}$$

最后，将标准差和均值的比值作为多样性指标，记为 $Diversity$，如式(23)所示，该值越大说明度量方法越好。如果 σ_{dist} 较大，说明语义向量间的距离差异较大，由于不同的度量方法得到的距离的取值范围不同，导致了距离的量纲不同，标准差大未必能表示多样性好，应同时比较均值和标准差才能表明 $\boldsymbol{S}$ 中的数据具有多样性。

$$Diversity=\sigma_{dist}/\mu_{dist} \tag{23}$$

4.2.2　语义性指标

文本中的词语用语义向量表达，则词语间的语义关系可用语义向量间的某种关系表达。然而，自然语言中的语义关系非常复杂，本文研究的语义向量差异性仅适合表达语义相似或相关关系，因此针对本文的实验数据 DataSet1，选取了两个出现次数较高、语义相似或相关、在渔业领域中较为典型的词语：“水产”和“捕捞”，其在语料中出现的次数分别为 1155 和 711。

语义相似或相关的词语间的语义向量距离应该较“小”，但很难给出具体的“大”或“小”的界限，因此本文在实验中给出与每个词语距离最小的 5 个词语，然后从语言学角度评价这 5 个词语与给定的词语在语义上相似或相关的程度，进而对度量方法进行评价。

4.3　实验结果和分析

针对以上两个评价指标，本文分别在 DataSet1 和 DataSet2 两个数据集上对多种度量方法的多样性和语义性进行比较，以验证基于非对称多值特征杰卡德差异性度量方法的度量效果。

4.3.1　基于连续数据和离散数据的不同度量方法的多样性比较

实际应用中有很多针对连续数据的度量方法，本文选取了具有代表性的 3 种方法：向量空间中的 L_p 范数度量方法 P2_Dist、内积空间中的内积距离度量方法 Inner_Dist 和考虑统计信息的相关距离度量方法 Corr_Dist。首先将这 3 种方法与 J_Dist 进行比较，结果如图 6 所示。

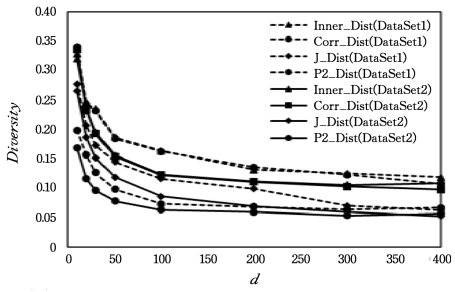

图 6　连续和离散数据的度量方法的多样性比较

Fig. 6　Diversity comparison results of measurement methods for continuous and discrete data

从图 6 可得出，DataSet2 上的多样性略优于 DataSet1，但随维度 d 的变化其趋势是相同的。当语义向量的维度 d 增加时，多样性 $Diversity$ 的总体趋势变差；具体地，在维度 $d=10$ 时，J_Dist 在 DataSet1 上的 $Diversity=0.2654$，J_Dist 在

DataSet2 上的 $Diversity=0.2772$，都比 P2_Dist 好，但都差于 Inner_Dist 和 Corr_Dist；Corr_Dist 的多样性最优，在 DataSet1 上的多样性 $Diversity=0.3356$，在 DataSet2 上的多样性 $Diversity=0.3396$。这是因为 J_Dist 中离散化的粒度过粗，使得不同语义向量的维度之间的区分度变小，最终导致语义向量的差异性度量结果趋于集中。

4.3.2 JM_Dist 度量方法的多样性实验

(1)JM_Dist 方法的参数变化实验

本实验中，3.3 节步骤 2 中的重要性函数取初等函数中的反比例函数，即 $f(freq_{j,z})=1/freq_{j,z}$；步骤 3 中的 $\boldsymbol{T}^f$ 取值不同时，度量结果会不同，将 $\boldsymbol{T}^f$ 作为 JM_Dist 的第一个参数。另外，式(1)中 c 的取值对度量结果有直接影响，将其作为 JM_Dist 的第二个参数。经实验验证，在 DataSet1 数据集上，不同维度的语义向量的度量方法的差异性随参数 $\boldsymbol{T}^f$ 的变化情况类似，随 c 的变化情况也类似；在 DataSet2 数据集上的差异性效果随参数的变化情况与 DataSet1 上的非常相近，因此本文仅给出 DataSet1 上 10 维语义向量的多样性随 $\boldsymbol{T}^f$ 和 c 的变化情况，结果如图 7 所示。

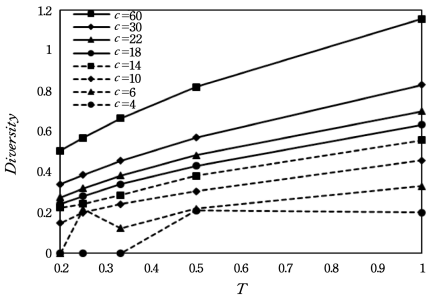

图 7 JM_Dist 随参数 $\boldsymbol{T}^f$ 和 c 变化的多样性

Fig. 7 Diversity of JM_Dist varies with parameters $\boldsymbol{T}^f$ and c

根据图 7，针对参数 c 得出两点结论：1)当参数 c 增大时，多样性提高，但提升越来越缓慢。例如 $\boldsymbol{T}^f=(1.0,\cdots,1.0)$ 时，在 DataSet1 上的 $Diversity$ 随 c 的变化情况如图 8 所示。

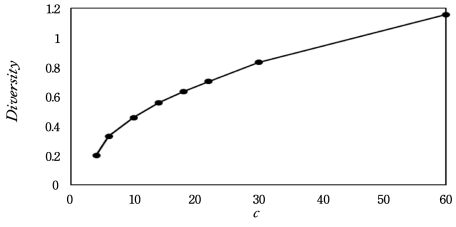

图 8 JM_Dist 随参数 c 变化的多样性

Fig. 8 Diversity of JM_Dist varies with parameters c

当 $c=10$ 时，$Diversity=0.4583$；当 $c=14$ 时，$Diversity=0.5584$；当 $c=18$ 时，$Diversity=0.6353$，即多样性提升的趋势变缓。2)$c=4$ 和 $c=6$ 时，多样性结果不稳定，从 $c=10$ 开始趋于稳定。所以在实际应用中，若提高参数 c 的值则可提高多样性，但随着 c 的增大，计算效率下降，因此建议 c 取值不宜过大。根据图 7，针对参数 $\boldsymbol{T}^f$ 得出，当参数 $\boldsymbol{T}^f$ 增大时，除了 $c=6$ 时略有不稳定变化外，其他 $Diversity$ 均降低，$\boldsymbol{T}^f=(1.0,\cdots,1.0)$ 时 $Diversity$ 最优，若不做额外说明，度量方法 JM_Dist 的 $\boldsymbol{T}^f$ 取值均为 $\boldsymbol{T}^f=(1.0,\cdots,1.0)$。

(2)JM_Dist 方法在不同维度时的多样性实验

根据 JM_Dist 方法的参数变化实验的结论，在 DataSet1 和 DataSet2 数据集上，JM_Dist 均在 $c\geqslant 14$ 时多样性较优，因此本节取 $c=14$ 和 $c=60$。经实验验证，DataSet1 和 DataSet2 上不同度量方法的多样性指标非常接近，本文仅给出 DataSet1 上不同维度的语义向量的 JM_Dist 度量方法与连续数据的度量方法的多样性比较结果，如图 9 所示。

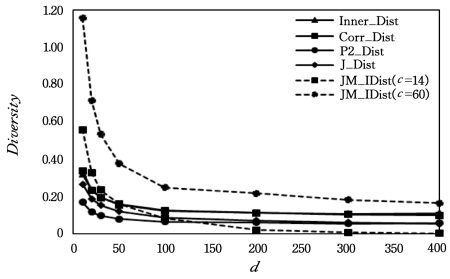

图 9 JM_Dist 方法与其他度量方法的多样性比较

Fig. 9 Diversity comparison results between JM_Dist and other measurement methods

根据图 9 得出两点结论：1)随着语义向量维度 d 的增加，所有度量方法的多样性都降低；2)整体上，当 $c\geqslant 14$ 时，JM_Dist 方法的多样性优于其他度量方法，证实了本文提出的方法缓解了“度量集中”问题。具体地，当 JM_Dist 方法中 $c=14$ 且语义向量维度 $d<50$ 时，JM_Dist 方法的多样性优于其他方法；当 $d\geqslant 50$ 时，JM_Dist 方法的多样性开始差于 Corr_Dist 方法，甚至差于其他所有的方法。提高参数 c 的值，当 $c=60$ 时，在 10～400 维语义向量下 JM_Dist 方法的多样性都明显优于 Corr_Dist 等其他方法。在 10 维语义向量下，JM_Dist($c=60$)方法的 $Diversity=1.1571$，连续数据中多样性最优的 Corr_Dist 方法的 $Diversity=0.3359$，提高了 0.8251，增幅较大。

基于以上实验建议参数 c 的取值为 $c\geqslant 14$，$\boldsymbol{T}^f=(1.0,\cdots,1.0)$，此时 JM_Dist 多样性最优，度量方法的多样性指标的排序为：JM_Dist($c=60$)＞JM_Dist($c=14$)＞Corr_Dist＞Inner_Dist＞J_Dist＞P2_Dist。

4.3.3 语义性评价

经实验验证，DataSet1 和 DataSet2 上的度量方法的语义性评价结果类似，因此本节仅给出 DataSet1 上的 JM_Dist，Corr_Dist，Inner_Dist，J_Dist 和 P2_Dist 方法的语义性评价结果，如表 2 所列。表 2 中所列词语与“水产”或“捕捞”的语义相似或相关。由表 2 和图 9 的实验结果得出两点结论：1)在基于连续数据的度量方法中，Inner_Dist 方法在多样性和语义性的综合效果方面最优。P2_Dist 方法的语义性较好，但其多样性差；Corr_Dist 的多样性较优，但语义性差；Inner_Dist 的语义性较优，但其多样性略差于 Corr_Dist，因此在连续数据的度量方法中 Inner_Dist 总体上最优。2)所有方法中，JM_Dist 在 $c=60$ 时综合效果最优。J_Dist 方法的多样性和语义性都较差；JM_Dist 在 $c=14$ 时多样性优于所有的连

续数据的度量方法，其语义性较优；JM_Dist 在 $c=60$ 的语义性与 $c=14$ 时差别不大，但其多样性优于 $c=14$ 时的效果，因此在所有方法中该方法总体上最优，从而验证了本文提出的方法的有效性。若读者使用度量方法的应用场景对时间效率要求不高，可适当增大参数 c 的值，以进一步改善多样性效果。

表 2　DataSet1 上的语义性评价结果

Table 2　Semantic evaluation results on DataSet1

方法	d	词语(top5)	
		水产	捕捞
JM_Dist ($c=60$)	10	细胞膜，15 天，食鱼，验性，采收	苗种，变革，船式，清理，编绳
	20	畜类，烤鳗，抵达，船内，北极苗鱼	生产，缠络，珊瑚鱼，鱼病，右旋
	30	水獭，掀起，气象学，冷藏车，吸热	鱼病，白蝶贝，捻度，盛开，产出
	50	**产自，渔业，湘云鲫，水域，淡水河**	**掠夺，渔捞，水域，绳，藻**
	100	**生产，夏季，资源，紫菜，水域**	**拔起，人工，淡水河，采苗，船式**
	200	**养殖，渔业，咸度，捕捞，船舶**	**切开，采苗，捕获，渔捞，资源**
	300	**渔业，资源，潮水，网箱，海水**	**采苗，海水，浅表，捕获，水产**
	400	**亚寒带，捕捞，动物，养殖，海洋**	**领海基线，水产，捕获，养殖，近距离**
JM_Dist ($c=14$)	10	机型，裸鲤，耙齿，验性，曾对	续航，白鳝，清理，嵊山，编绳
	20	亚寒带，烤鳗，流水，壳，该类	探头，养殖，西风带，操纵性，不定期
	30	反水，南流，鲔，水产品，出船坞	饲料，增殖，感染，网目，续航
	50	**分布图，锥状，渔业，湘云鲫，水域**	**补助，供应船，产出，水产，近海**
	100	**航海日志，夏季，资源，生产，水域**	**花鲈，捕获，锚固，无线电通，资源**
	200	**养殖，渔业，咸度，捕捞，船舶**	**飞鱼，种群，杂合，港湾，海藻**
	300	**渔业，船底，潮水，网箱，环境**	**融冰，海藻，养殖，生产，捕获**
	400	**团，水产品，动物，养殖，海洋**	**水产，捕获，港湾，冷空气，船底**
Corr_Dist	10	类经，跃升，原液，舫，工资	传世，基围虾，中产，大堡礁，燃油
	20	内在，牛舌，白鲑，促排卵，甲藻门	内在，交通事故，死疫苗，少数分，缓步
	30	间距，铁细菌，表面上，糕，胶袋	成败，蓝边，光性，麻痹性，扎在
	50	强冷空气，化上，衣装，信息化，一门	5 万，小包装，幕墙，中长，患
	100	棉，污染源，泥螺，鱼粉，全国人	届，代表大会，库，免疫记忆，泥螺
	200	民，全国人，解冻，常务委员，1995 年	常务委员，全国人，民，解冻，泊位
	300	全国人，污染源，泥螺，解冻，民	全国人，常务委员，污染源，民，鱼粉
	400	全国人，民，污染源，代表大会，种群	民，解冻，常务，委员，泥螺
Inner_Dist	10	胶州湾，分装，弗罗曼，沿途，负责	对照组，试验，系统结构，表层水，保留
	20	**渔业，捕捞，资源，养殖，海洋**	**水产，养殖，渔业，资源，内面**
	30	**养殖，水产品，渔业，捕捞，成体**	**养殖，水产，渔业，外延，水产品**
	50	**养殖，资源，捕捞，渔业，水产品**	**水产，养殖，资源，渔业，水域**
	100	**养殖，渔业，资源，捕捞，生产**	**水产，资源，养殖，渔业，水域**
	200	养殖，渔业，资源，生产，水产品	水产，养殖，资源，渔业，生产
	300	养殖，渔业，资源，水产品，生产	水产，资源，渔业，养殖，水域
	400	养殖，渔业，资源，水产品，动物	水产，资源，养殖，渔业，水域
J_Dist	10	鳞片状，温水，喷头，28 个，网眼	有序，厦门，后裔，鳞片状，温水
	20	箭，海礁，坑塘，起吊，有刺	借此，鳍足，释放出，学法，起吊
	30	活泼，耐光，加厚，陀螺，椎	坐在，字母，等压，之也，鳍足
	50	武昌鱼，基金会，氯离子，一门，雄鱼	中长，基金会，可氧化，雀鳝，雄鱼
	100	延伸线，过了，稍有不慎，棉，49 对	合时，污染源，结有，鱼粉，高位
	200	解冻，民，污染源，一顶，泊位	解冻，代表大会，民，控制点，一顶
	300	全国人，民，污染源，洄游，种群	洄游，污染源，蓄积，全国人，泊位
	400	民，泥螺，解冻，污染源，监测	全国人，届，污染源，解冻，泥螺
P2_Dist	10	胶州湾，分装，弗罗曼，沿途，负责	对照组，试验，系统结构，表层水，保留
	20	渔业，捕捞，资源，养殖，海洋	水产，养殖，渔业，资源，内面
	30	**养殖，水产品，渔业，捕捞，成体**	**养殖，水产，渔业，外延，水产品**
	50	**养殖，资源，捕捞，渔业，水产品**	**水产，养殖，资源，渔业，水域**
	100	**养殖，渔业，资源，捕捞，生产**	**水产，资源，养殖，渔业，水域**
	200	养殖，渔业，资源，生产，水产品	水产，养殖，资源，渔业，生产
	300	养殖，渔业，资源，水产品，生产	水产，资源，渔业，养殖，水域

结束语　本文从离散数据的角度给出一种解决“度量集中”问题的高维语义向量的度量方法，从多样性和语义性两个角度对该方法进行了评价。该方法较目前文献中的度量方法有大幅提高，为自然语言处理的诸多应用中的度量方法提供了一种新的解决思路。

但由于语义学上的语义关系复杂，导致语义距离复杂，单一的距离无法很好地表达复杂的语义距离，能否结合多种度量方法给出一种线性或非线性加权的综合度量方法，以更准确地表达语义距离，是下一步要探讨的问题之一。

语义向量是根据语言模型训练神经网络而得到的，通过不同的神经网络结构和代价函数会得到不同的语义向量，度量效果是否与二者有关是需要进一步研究的问题之二。

本文提出的度量方法仅针对自然语言处理领域的语言模型中的语义向量，而在图像和语音领域的应用中也需要度量方法，该度量方法是否适用或者有无其他更好的度量方法，是进一步要研究的问题之三。

参 考 文 献

[1] 中文信息处理发展报告[EB/OL].[2017-4-11]. http://www.cipsc.org.cn/download.php? file=cips2016.pdf.

[2] PACCANARO A,HINTO G E. Learning distributed representations of concepts using linear relational embedding[J]. IEEE Transactions on Knowledge & Data Engineering,2001,13(2):232-244.

[3] BENGIO Y,SCHWENK H,SENÉCAL J,et al. Neural Probabilistic Language Models[J]. Journal of Machine Learning Research,2001,3(6):1137-1155.

[4] FENG Y H,YU H,SUN G,et al. Domain-specific Terminology Recognition Method Based on Word Embedding and CRF[J]. Journal of Computer Applications,2016,36(11):3146-3151. (in Chinese)
冯艳红,于红,孙庚,等. 基于词向量和 CRF 的领域术语识别方法[J]. 计算机应用,2016,36(11):3146-3151.

[5] YAN J,LIU W F,LIN H F. Music Recommendation Study Based on Tags Multi-Space[J]. Journal of Chinese Information Processing,2014,28(4):117-122. (in Chinese)
闫俊,刘文飞,林鸿飞. 基于标签混合语义空间的音乐推荐方法研究[J]. 中文信息学报,2014,28(4):117-122.

[6] MIKOLOV T,CHEN K,CORRADO G,et al. Efficient Estimation of Word Representations in Vector Space[J]. Computer Science,arXiv:1301.378lv3.

[7] MIKOLOV T,SUTSKEVER I,CEHN K,et al. Distributed representations of words and phrases and their compositionality[J]. Advances in Neural Information Processing Systems,2013,26:3111-3119.

[8] BUHLMANN P,VAN DE GEER S. Statistics for High-Dimensional Data[M]. Springer-Verlag Berlin Heidelberg,2011.

[9] BELLMAN R. Adaptive Control Process:A Guide Tour[M]. Princeton University Press,Princeton,New Jersey,1961.

[10] FUKUNAGA K. Introduction to Statistical Pattern Recognition (2nd ed)[M]. New York:Academicpress,1972.

[11] LEDOUX M. The concentration of measure phenomenon[J]. Mathematical Surveys and Monographs,2001,89:94-124.

[12] HE L,CAI Y C,YANG Z. Researches on Similarity Measurement of High Dimensional Data[J]. Computer Science,2010,37(5):155-156. (in Chinese)
贺玲,蔡益朝,杨征. 高维数据的相似性度量研究[J]. 计算机科学,2010,37(5):155-156.

[13] HE J R,DING L X,HU Q H,et al. Properties of High-dimensional Data Space and Metric Choice[J]. Computer Science,2014,41(3):212-217. (in Chinese)
何进荣,丁立新,胡庆辉,等. 高维数据空间的性质及度量选择[J]. 计算机科学,2014,41(3):212-217.

[14] CHEN S G,ZHANG D Q. Experimental Comparisons of Semi-Supervised Dimensional Reduction Methods[J]. Journal of Software,2011,22(1):28-43. (in Chinese)
陈诗国,张道强. 半监督降维方法的实验比较[J]. 软件学报,2011,22(1):28-43.

[15] FENG L,LIU S L,ZHANG J,et al. Robust Activation Function of Extreme Learning Machine and Linea Dimensionality Reduction in High-Dimensional Data[J]. Journal of Computer Research and Development,2014,51(6):1331-1340. (in Chinese)
冯林,刘胜蓝,张晶,等. 高维数据中鲁棒激活函数的极端学习机及线性降维[J]. 计算机研究与发展,2014,51(6):1331-1340.

[16] LAI S W. Word and Document Embedding Based on Neural Network Approaches[D]. Beijing:University of Chinese Academy of Sciences,2016:27-39. (in Chinese)
来斯惟. 基于神经网络的词和文档语义向量表示方法研究[D]. 北京:中国科学院大学自动化研究所,2016:27-39.

[17] JACCARD P. Etude de la distribution florale dans une portion des Alpes et du Jura[J]. Bulletin De La Societe Vaudoise Des Sciences Naturelles,1901,37(142):547-579.

[18] Jaccard index[EB/OL].[2017-4-29]. https://en.wikipedia.org/wiki/Jaccard_index#cite_note-1.

[19] SAMANTHULA B K,JIANG W. Secure Multiset Intersection Cardinality and its Application to Jaccard Coefficient[J]. IEEE Transactions on Dependable & Secure Computing,2016,13(5):1.

[20] CHENG Y,WANG S T. A Multiple Alternative Clusterings Mining Algorithm Using Locality Preserving Projections[J]. CAAI Transactions on Intelligent Systems,2016,11(5):600-607. (in Chinese)
程旸,王士同. 基于局部保留投影的多可选聚类发掘算法[J]. 智能系统学报,2016,11(5):600-607.

[21] LIAO B,ZHANG T,YU J,et al. Efficiency Optimization of Jaccard's Similarity Coefficient Based on Two Dimensional Partition[J]. Computer Science,2017,44(1):219-225. (in Chinese)
廖彬,张陶,于炯,等. 基于二维划分的杰卡德相似系数批量计算效率优化[J]. 计算机科学,2017,44(1):219-225.

[22] TANIMOTO T T. An Elementary Mathematical theory of Classification and Prediction[R]. Internal IBM Technical Report,1957.

[23] ROGERS ,TANIMOTO D J,TAFFEE T. A Computer Program for Classifying Plants[J]. Science,1960,132(3434):1115-1118.

[24] 潘迎捷. 水产辞典[M]. 上海:上海辞书出版社,2007.

[25] 搜狗全网新闻数据(SogouCA)[EB/OL].[2017-02-14]. http://www.sogou.com/labs/dl/ca.html.