基于互异特征向量的 ERP 重复物料记录识别方法研究*

宗威[1]，林松涛[1]，蒿恒[1]，吴锋[2]
（1. 西安电子科技大学 经济与管理学院，西安 710071；2. 西安交通大学 管理学院，西安 710049）

摘　要　重复物料记录是 ERP（enterprise resource planning，企业资源计划）系统实施初期普遍存在的数据质量关键问题。自然语言的复杂性与物料命名的多样性增加了重复物料记录识别的难度。传统重复记录识别方法大都基于字符匹配或向量匹配，忽略字符间的语义相似性，很难有效解决重复物料记录识别问题。通过研究物料实体的命名机理，首先对物料名称进行反转、排序、倒排索引操作，构建基于概率神经网络（probabilistic neural network，PNN）的互异特征向量模型，通过学习互异特征向量之间的语义相似性实现 ERP 重复物料记录识别。

关键词　互异特征向量，重复物料记录，语义相似性，数据质量，ERP

中图分类号　C931.6

1　引言

随着全球经济一体化及竞争环境的日益激烈，越来越多的企业已经实施 ERP 系统，ERP 系统的实施已经成为最近几十年影响最大、最普遍的企业变革之一[1]。然而令人遗憾的是，尽管企业在实施 ERP 系统过程中投入了大量资源，但并没有获得预期的效果，ERP 实施的失败率和实施难度依然很高。Hung 等学者指出，即使是在咨询公司及行业最佳实践者的帮助下，ERP 系统的实施失败率依然超过 50%[2]。2014 年 Panorama 咨询公司发布的一项调查报告显示，ERP 系统的预算超支比例、实施超期比例、少于预期收益一半的比例分别达到了 54%、72%、66%[3]。

在影响 ERP 实施效果的众多关键因素中，数据质量已经成为备受企业关注的影响因素之一，数据质量是决定 ERP 系统实施成败的关键因素[4~6]。数据是 ERP 系统的灵魂，“三分软件、七分组织、十二分数据”已经成为业界公认的真理，充分体现了数据的重要性。ERP 系统的有效运作以高质量的数据为前提，数据价值得以充分、有效发挥的前提是必须保证 ERP 系统的数据质量，数据质量是实现企业决策有效性的关键环节[7]。对于 ERP 系统而言，物料数据是制造企业信息系统中最重要的基础主数据之一，它包含了对企业所采购、生产和存储在库存中物料的详尽描述，描述了有关物料的所有信息。物料数据是以 ERP 为核心的企业信息系统的运行基础，也是其他运营数据［如物料需求计划、主生产计划、BOM（bill of material，物料清单）等数据］参照的关键依据[8]。因此，保证物料数据的高质量

* 基金项目：国家重点研发计划资助项目（2018YFB1703003，2018YFB1703000）、中国博士后科学基金资助项目（2017M623130）、中央高校基本科研业务费专项资金资助项目（JB190609）。

通信作者：宗威，西安电子科技大学经济与管理学院讲师，硕士生导师。E-mail：weizong@xidian.edu.cn。

是实现 ERP 系统正常运行的重要前提。

然而，在 ERP 系统实施过程中，尤其是在 ERP 系统实施初期，重复物料记录已经成为目前企业，特别是制造企业在实施 ERP 过程中普遍面临的数据质量问题，严重影响了数据的唯一性与准确性。一方面，在 ERP 系统实施初期，企业和 ERP 系统之间仍处于磨合阶段，物料编码规则与观念没有深入人心，企业的编码、设计、生产及仓储等各职能部门仍存在各自为政的现象，使得各部门在增加、删除、修改物料数据时沟通不够及时、顺畅，造成了重复物料记录的存在；另一方面，由于制造企业的产品设计人员众多且分属不同的产品设计部门，特别在 ERP 系统实施初期，不同部门的产品设计人员在物料命名习惯、书写格式、表达方式等方面还没有达成统一的共识，经常存在很大差异，ERP 系统中极易出现一物多码的重复物料数据。重复的物料数据不仅会影响物料核算及采购决策的准确性，甚至会给整个生产过程带来混乱。因此，如何有效识别 ERP 系统中重复的物料记录是一项十分必要且重要的研究工作。

2 相关研究述评

重复记录识别，又称为实体识别，是在数据库中识别哪些记录表示现实世界同一实体的过程[9, 10]。重复记录的识别和检测是学术界和业界普遍关心的问题，已经引起了数据库、信息系统及其他相关领域学者的研究兴趣，对该问题的研究取得了丰富的成果。

重复记录识别过程包括两个阶段：首先，比较记录对中同一属性之间的字符串相似性；其次，通过综合每对属性的字符串相似性，判断两条记录的整体相似性。基于文本相似度函数的字符串匹配方法是测度记录相似性的常用方法，即根据两条记录对应属性的字符匹配程度来判断两条记录的相似性[11, 12]。编辑距离是比较记录字段相似性的经典算法之一，通过计算一个字符串转换成另一个字符串所需的最少编辑操作（插入、删除、替换）次数判断字符串之间的相似性[13]。Q-grams 方法是另一种基于字符匹配思想的、通过比较两字符串共有的子字符串来判断字符串的相似性的方法[14]。Cosine 相似度，是一种基于向量空间的字符串匹配算法。它首先将字符串转换成向量的形式，通过计算向量的夹角余弦值判断字符串之间的相似性。TF-IDF（term frequency-inverse document frequency）是一种常用的文本加权技术，Cosine 方法经常与 TF-IDF 一起使用，计算字符串的相似性。向量空间模型（vector space model，VSM）是信息检索和文本挖掘领域常用的文本相似度计算方法，通过字符出现的频次与权重计算文本之间的相似性[15]。

通过对上述相关文献的分析总结发现，当前主流的重复记录识别算法大都基于如下假设：无论是由于输入错误还是不同的表达等原因，同一实体的属性值在表象上相差不大。例如，“Jone Doe”与“Jonn Doe”，虽然字符串表象上有细微差异，但有可能表示的是同一个人。然而，这个假设在识别 ERP 重复物料记录时并不成立。对于物料名称来说，即使两条物料记录在命名及表达上相差甚远，但它们仍然可能代表的是同一种物料；或者两条物料记录在命名及表达上极为相近，但有可能代表的是完全不同的两种物料。例如，“无油轴承”与“自润滑轴承”，不同的名字描述的却是同一种物料；又如，“无油轴承”与“有油轴承”，相近的物料名称，描述的却是两种不同的物料。也就是说，在识别重复物料记录方面，不仅仅要考虑字符串本身，还要考虑字符之间的语义相似性。

尽管文本挖掘过程中考虑了文本之间的语义相似性，但语义相似性的衡量与计算通常借助于 HowNet 或 WordNet 等通用的外部知识库。对于物料数据来说，HowNet[16]和 WordNet[17]词典系统中并不包括描述物料数据的一些特殊及专业词汇。另外，描述物料记录的关键词数量庞大，不可能事先标

记出所有关键词的语义相似度。特别是各个企业根据其自身环境与发展需要而自主设立的描述规则与描述方式具有很大的灵活性和不规范性，这些现实的客观因素都为有效识别 ERP 系统中的重复物料记录带来了难度和挑战。因此，为了弥补传统字符串匹配算法在语义相似性方面的缺陷，以及外部通用知识库在解决新词汇与特殊词汇方面的不足，本文基于 ERP 物料实体自身的特征信息，提出了一种基于互异特征向量的 ERP 重复物料记录识别方法，其基本思想是：两条物料记录之间的差异性是由物料记录名称之间的差异词决定的。因此，本文运用有监督学习的机器学习方法——PNN 来学习物料名称之间差异词语义相似性的流程与方法，通过学习与识别差异词之间的语义相似性来判断并预测表达不同的两种物料指代的是否为同一种物料实体，实现对 ERP 重复物料记录的有效识别。

3 基于互异特征向量的 ERP 重复物料记录识别流程与方法

根据物料名称的命名特点，构建了基于互异特征向量的 ERP 重复物料记录识别流程与方法。整体的流程与方法由数据预处理、索引和分块、训练和学习三部分构成。下面以“R_1=无油轴承，R_2=方形固定支撑块，R_3=深沟球轴承，R_4=自润滑轴承，R_5=方形支撑块”这 5 条物料记录为例，详细介绍重复物料记录的具体识别过程，如图 1 所示。

3.1 数据预处理

数据预处理由数据倒置和排序两部分构成，是后续数据处理的基础。通过分析物料名称命名特点发现，描述物料本质的名词一般位于物料名称的结尾处。以“轴承”物料为例，深沟球轴承、无油轴承、自润滑轴承等物料的本质是“轴承”，它位于描述性语言“深沟球”“无油”“自润滑”等词语之后。根据物料名称的这个特点，首先将物料名称进行倒置处理，然后将倒置的物料名称记录按照字母顺序升序排列（见图 1 中“数据预处理”部分示例）。若采用传统的排序方法，描述性语言首字母字典位置的差异，如“深沟球轴承”（首字母为 S）和“自润滑轴承”（首字母为 Z），会使得本质相同的两种物料在排序位置上相差很远，不利于记录的聚类。然而，当进行倒置、排序处理之后，两者变为“承轴球沟深”（首字母为 C）和“承轴滑润自”（首字母为 C），这样属于同种类别的物料就很容易被排列在一起，方便物料记录的聚类。

3.2 索引和分块

传统的重复记录识别方法是比较所有记录构成的记录对，然而对于大型数据库来说，不仅费时，甚至是不可行的。为了减少记录比较次数，提高重复物料识别效率，本文在倒置并排序后的物料名称上运用倒排索引的方法，将数据库中的记录根据物料所属类别划分成相互独立的记录块（block），只比较同一个记录块中的记录所形成的记录对，进行重复物料记录的识别。其中代表物料类别的词语作为倒排索引的关键字，如“齿轮”“轴承”“支撑块”等，这样具有相同关键词的物料记录被聚集在同一个记录块中。例如，所有描述轴承的物料记录经过倒排索引后被聚在“轴承”中，而所有描述支撑块的物料记录则被聚在“支撑块”中。重复物料记录的识别分别在“轴承”和“支撑块”中进行，无须比较“轴承”和“支撑块”之间的记录，大大提高了重复物料记录的识别效率。索引和分块的具体操作过程见图 1 中“索引和分块”部分示例。

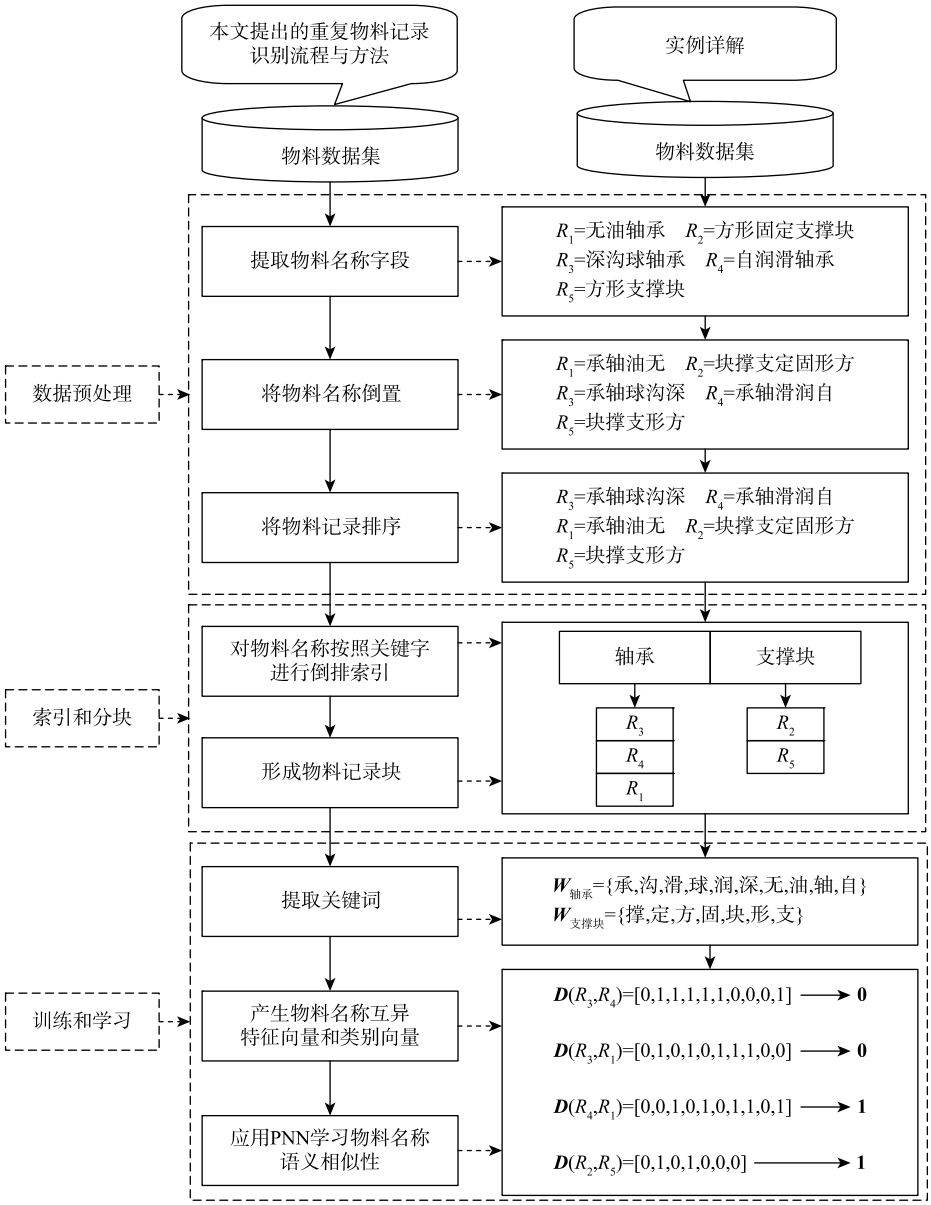

图 1　基于互异特征向量的 ERP 重复物料记录识别流程与方法

3.3　构建互异特征向量模型

本文主要研究的是如何通过物料名称识别重复物料记录，物料名称描述的差异是决定记录重复与否的关键，因此，可以通过学习物料名称之间的差异词来判断字符之间的语义相似性，进而判断是否为重复记录。首先，不重复地提取一个记录块中的关键字集合（不包括“的”“和”“并”“是”等停用词），$\boldsymbol{W}=\{\text{word}_1,\text{word}_2,\cdots,\text{word}_n\}$，并按照字母顺序升序排列（如图 1 中“提取关键词”部分示例）。其次，将相互比较的记录对表示成互异特征向量的形式，$\boldsymbol{D}(R_s,R_t)=[M_1,M_2,\cdots,M_n]$，其中 R_s 和 R_t 表示相比较的两个物料名称。M_i 是一个 0-1 变量，若 $M_i=0$，则 word_i 是两个物料名称的共有词或者两个名称中都不包括的词；若 $M_i=1$，则 word_i 是两个物料名称的差异词，即 word_i 只在 R_s 中或者只在 R_t 中。物料类别向量由 $\boldsymbol{C}$ 表示，若 $\boldsymbol{C}=\boldsymbol{0}$，代表两个物料名称语义表达上存在差异，如果 $\boldsymbol{C}=\boldsymbol{1}$ 则代表物

料名称语义表达相同（如图 1 中“训练和学习”部分示例）。

3.4 学习物料名称之间的语义相似性

PNN 是基于最小风险贝叶斯决策规则的模式分类方法，由输入层、模式层、累加层和输出层构成，PNN 的结构见图 2。

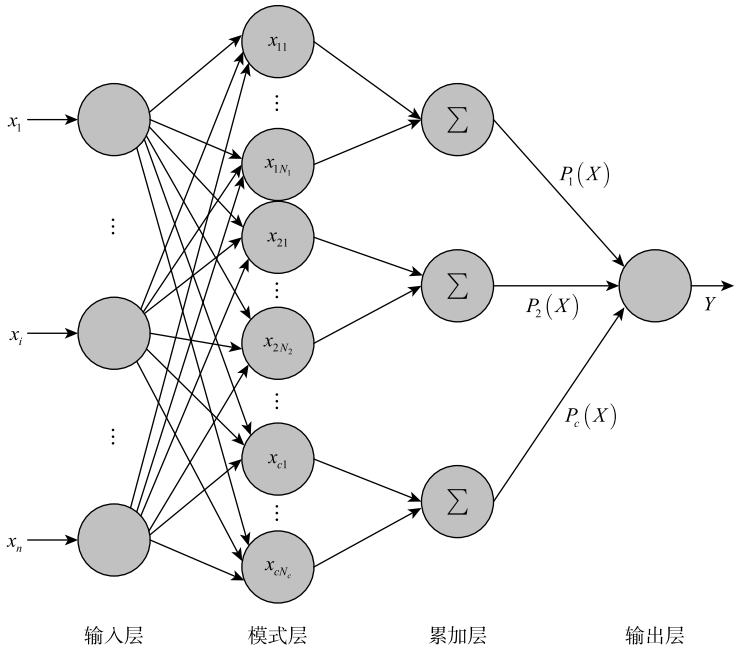

图 2 PNN 结构图

输入层中，$\boldsymbol{X}=(x_1,x_2,\cdots,x_n)$表示 n 维的输入向量；模式层中的 N_c 表示第 c 类中的训练样本数量，x_{ij} 表示属于第 i 类的第 j 个训练样本，模式层神经元的输出表示 $\boldsymbol{X}$ 与 x_{ij} 之间的似然值。累加层共有 c 个神经元，代表 c 个不同的类别，累加层中的每一个神经元对来自同一个类别的模式层神经元的输出进行加总求和，表示 $\boldsymbol{X}$ 属于第 i 类的似然概率 $P_i(X)$。最后，输出层神经元通过比较 $\boldsymbol{X}$ 属于各个类别的 $P_i(X)$ 值，确定 $\boldsymbol{X}$ 最终所属的类别。与其他传统的 BP（back propagation，反向传播）神经网络相比，PNN 具有训练速度快、样本追加方便、在训练之前无须确定隐含层神经元的数量，即使在噪声数据中仍能得到比较稳健的分类结果等特点，在语音识别、信号处理、医疗诊断等很多领域得到了广泛应用，能够很好地解决分类问题[18]。因此，本文运用 PNN 方法学习物料名称之间的语义相似性，对重复与非重复的物料记录进行分类处理。应用 PNN 机器学习方法识别重复物料记录分为训练和测试两个阶段。在训练阶段，将已经转化的物料名称差异词向量 $\boldsymbol{D}(R_s,R_t)$ 和类别向量 $\boldsymbol{C}$ 分别作为 PNN 的输入和输出，训练 PNN 结构。在测试阶段，将待测试的物料名称差异词向量输入 PNN，经过训练的 PNN 根据学习的结果，判断物料名称的语义相似性，并输出 **0** 或 **1** 的类别向量。

4 实验及结果分析

为了验证本文提出的重复物料记录识别流程和方法的有效性，采用来自实际企业的物料数据作为

实验对象。该企业位于陕西省，是生产汽车热交换器的专业厂家，也是中国西北地区唯一的汽车散热器专业生产企业，有着 30 余年的生产历史。该企业自 2011 年 6 月开始实施 ERP 系统以来，由于物料名称表达和描述的不一致，ERP 中产生了大量的重复物料记录，严重影响了 ERP 系统的实施效果。此次实验共收集到 2 678 条物料记录（训练样本 1 646 条，测试样本 1 032 条），经过初步的数据预处理，剔除不完整及重复数据后，最终参与实验验证的数据共 2 007 条（训练样本 1 209 条，测试样本 798 条），训练样本和测试样本中重复物料记录的标注工作由该企业的专业物料管理人员完成，实验样本数据分布如表 1 所示。其中，每一个物料类别中包含多种物料，每一种物料描述不同导致在系统中会产生重复物料记录。以表 1 中的轴承类物料为例，在收集的轴承类物料数据中，包含无油轴承、滚动轴承等多种轴承类物料记录，针对无油轴承而言，还有自润滑轴承这一等价的描述方式，因此产生了重复物料记录。

表 1　实验样本数据分布情况表

序号	物料数据类别	样本数量	训练样本数量	测试样本数量
1	轴承类（如无油轴承、滚动轴承等）	650	383	267
2	模块类（如数字输出/输入模块、拓展模块等）	91	54	37
3	变频器类（如电压变频器、电流逆变器等）	66	49	17
4	编码器类（如光电编码器、绝对值编码器等）	120	69	51
5	开关类（如电子压力开关、机械压力开关等）	210	116	94
6	板类（如防滑链板、支撑板、连接板等）	666	415	251
7	杆类（如支撑杆、导杆、推杆等）	84	51	33
8	连接块类（如固定连接块、导向块等）	120	72	48
合计		2 007	1 209	798

为了验证本文所提出方法的有效性，分别与基于字符的经典算法——编辑距离，以及基于向量空间的经典算法——VSM 相比较，并采用重复记录识别领域中常用的三个衡量指标评价方法的有效性，即精确率（precision）、召回率（recall）和 F_1 测度值[14]，计算公式分别如式（1）、式（2）、式（3）所示。

$$\text{precision}(P)=\frac{\text{TP}}{\text{TP}+\text{FP}} \tag{1}$$

$$\text{recall}(R)=\frac{\text{TP}}{\text{TP}+\text{FN}} \tag{2}$$

$$F_1=\frac{2\times P\times R}{P+R} \tag{3}$$

其中，TP 表示算法正确识别出的重复记录数量，（TP + FP）表示算法识别出的重复记录总数，（TP + FN）表示数据库中的重复记录总数。精确率表示预测为重复记录的那些数据中预测正确的数据个数。召回率则表示真实为重复记录的那些数据中预测正确的数据个数。F_1 测度值则是精确率和召回率的调和平均数。

为了更有助于理解公式所表达的数学含义，构造如表 2 所示的分类评价模型混淆矩阵。

表 2 分类评价模型混淆矩阵

混淆矩阵指标	预测为重复记录	预测为非重复记录
真实为重复记录	TP	FN
真实为非重复记录	FP	TN

另外，由于本文将重复记录识别作为分类问题处理，因此引入分类准确性作为评价方法有效性的又一个衡量指标，如式（4）和式（5）所示。其中，N 是测试样本总量，n_i 代表第 i 个测试样本，nc 是样本 n 所属的类别，classify（n）是本文方法识别出的 n 所属的类别。

$$\text{Classification Accuracy}(\text{CA}) = \frac{\sum_{i=1}^{N} \text{assess}(n_i)}{P+R}, \quad n_i \in N \tag{4}$$

$$\text{assess}(n) = \begin{cases} 1, & \text{classify}(n) = \text{nc} \\ 0, & \text{其他} \end{cases} \tag{5}$$

由于 PNN 算法是通过分析测试样本的分类结果来判断其有效性的，为了保证算法之间对比效果的科学性与合理性，参与测试的编辑距离法、VSM 方法在各个评价指标上的取值结果也是基于实验数据中的测试样本而获得的。基于 PNN 的重复物料识别方法与编辑距离、VSM 方法在四个指标上的效果比较如表 3 所示，此时相似度阈值设置为 0.7。从表 3 可以明显看出，基于 PNN 的重复物料识别方法在四个指标上的得分均高于其他两个方法，因为 PNN 方法考虑了物料名称之间的语义相似性，使得重复记录的识别结果更加准确、可靠。

表 3 重复物料记录识别方法效果比较

识别方法	分类准确性	精确率	召回率	F_1 测度值
基于 PNN 的方法	0.988 3	0.939 1	0.996 4	0.966 9
编辑距离法	0.835 2	0.555 6	0.217 4	0.313 0
VSM 方法	0.831 5	0.600 0	0.065 2	0.100 0

此外，为了验证方法的稳定性，本文进一步测试了相似度阈值的变化对重复物料记录识别结果的影响，即当相似度阈值变化时，识别方法在四个评价指标方面的变化情况，实验结果如表 4~表 7 所示。

表 4 阈值变化时三种方法分类准确性的变化趋势

分类准确性	相似度阈值			
	0.5	0.6	0.7	0.8
基于 PNN 方法的分类准确性	0.988 3	0.988 3	0.988 3	0.988 3
编辑距离法的分类准确性	0.797 8	0.846 4	0.835 2	0.835 2
VSM 方法的分类准确性	0.831 5	0.831 5	0.831 5	0.831 5

表 5 阈值变化时三种方法精确率的变化趋势

精确率	相似度阈值			
	0.5	0.6	0.7	0.8
基于 PNN 方法的精确率	0.939 1	0.939 1	0.939 1	0.939 1
编辑距离法的精确率	0.433 3	0.600 0	0.555 6	0.750 0
VSM 方法的精确率	0.545 5	0.571 4	0.600 0	0.666 7

表 6　阈值变化时三种方法召回率的变化趋势

召回率	相似度阈值			
	0.5	0.6	0.7	0.8
基于 PNN 方法的召回率	0.996 4	0.996 4	0.996 4	0.996 4
编辑距离法的召回率	0.565 2	0.326 1	0.217 4	0.065 2
VSM 方法的召回率	0.130 4	0.087 0	0.065 2	0.043 5

表 7　阈值变化时三种方法 F_1 测度值的变化趋势

F_1 测度值	相似度阈值			
	0.5	0.6	0.7	0.8
基于 PNN 方法的 F_1 测度值	0.966 9	0.966 9	0.966 9	0.966 9
编辑距离法的 F_1 测度值	0.491 0	0.422 3	0.310 3	0.120 0
VSM 方法的 F_1 测度值	0.210 5	0.151 0	0.100 0	0.081 7

基于上述实验验证数据，可以得出如下三点结论。

首先，当相似度阈值从 0.5 变化至 0.8 时，基于 PNN 的重复物料记录识别方法在分类准确性、精确率、召回率及 F_1 测度值四个指标方面没有变化，这是因为 PNN 方法主要是通过学习训练样本中物料名称之间的语义相似性来识别重复物料记录的，并不是通过与相似度阈值的比较来识别重复记录的，因此其识别效果并不受相似度阈值变化的影响；此外，由于 PNN 方法考虑了重复物料之间的语义相似性，其在各个评价指标的得分整体要高于编辑距离法和 VSM 方法。

其次，对于编辑距离法和 VSM 方法而言，当相似度阈值从 0.5 变化至 0.8 时，两种方法的精确率总体呈上升趋势。这是因为，当相似度阈值较小时，描述不同物料的物料名称会更容易被认为描述的是同一种物料，导致混淆矩阵中的 FP 值增加，使得识别的精确率［式（1）］降低；而随着相似度阈值的增加，一部分被误认为重复记录的非重复记录会被逐渐升高的相似度阈值剔除掉，导致混淆矩阵中的 FP 值变小，从而使得识别的精确率升高。

最后，对于编辑距离法和 VSM 方法而言，当相似度阈值从 0.5 变化至 0.8 时，两种方法的召回率呈下降趋势。这是因为，系统中真正重复的物料记录数量是固定的，也就是说对于召回率而言，公式（2）中分母的值是固定的，因此随着相似度阈值的增加，一部分描述相同物料的物料名称由于较高的相似度阈值而会更容易被认为描述的是不同的物料，导致混淆矩阵中的 FN 值增加，从而使得 TP 值变小，最终降低了识别的召回率。

通过上述实验验证结果可以发现，与只考虑字符之间表象特征的传统重复记录识别方法相比，基于 PNN 的重复物料记录识别方法由于考虑了记录之间的语义相似性，从而识别结果更加准确。因此，在 ERP 系统实施初期，企业物料管理人员可以根据图 1 所示的基于互异特征向量的 ERP 重复物料记录识别流程与方法对数据进行预处理后，进一步利用领域背景相关知识构造训练样本或者将系统中已知的重复物料记录作为训练样本构建互异特征向量，并通过 PNN 方法学习这些互异特征向量之间的语义相似性，进而对系统中未知的记录进行语义相似性识别，能够为最终判断是否为相似重复记录提供决策基础，提高识别效率和识别结果的准确性。

5 结论

数据质量是决定 ERP 实施成败的关键因素，重复物料记录是企业在实施 ERP 系统初期面临的普遍而严峻的数据质量问题。然而，物料数据的命名特点使得传统基于字符匹配或向量匹配的重复记录识别算法不能直接用于解决重复物料记录识别问题，因为这些传统算法只关注字符的表象特征，忽略了字符之间的语义相似性。物料名称命名的多样性及语义的复杂性，增加了重复记录识别的难度。本文以 ERP 系统中的物料记录为研究对象，首先，分析了物料记录的命名机理，提出了对物料数据进行倒置、排序、倒排索引等数据预处理方法；其次，根据重复物料实体之间的特征描述差异，构建互异特征向量模型，设计基于 PNN 的重复物料记录语义相似性识别算法。选取来自企业的真实数据进行数值分析，结果表明，本文提出的基于互异特征向量思想的 ERP 重复物料记录识别方法在精确率、召回率、分类准确性等评价指标方面的得分较高；且随着阈值的不断变化，基于互异特征向量思想的重复记录识别方法的识别效果较为稳定，说明其在识别重复物料记录方面是有效的。

本文研究的理论意义主要有：①本文刻画了 ERP 系统中物料数据这类具有短文本特性的实体命名机理，丰富了重复实体识别的理论研究；②本文提出了基于互异特征向量思想的 ERP 重复物料记录识别方法，为实体之间语义相似性的测度与识别提供了新的研究视角。

本文研究的实践意义主要有：①对于企业来说，该方法可以直接用于识别 ERP 系统中的重复物料记录，帮助企业及时获取并解决系统中重复记录的数据质量问题，有助于企业更好地管理物料数据，保证 ERP 系统中的数据质量与 ERP 系统的有效运行；②本文提出的流程与方法可以扩展至其他数据库中的重复记录识别，如商品数据库，通过识别重复的商品记录，可以进一步分析其价格走势、不同地区的销量等深层次信息。

参 考 文 献

[1] Zach O，Munkvold B E，Olsen D H. ERP system implementation in SMEs：exploring the influences of the SME context[J]. Enterprise Information Systems，2014，8（2）：309-335.
[2] Hung W H，Ho C F，Jou J J，et al. Relationship bonding for a better knowledge transfer climate：an ERP implementation research[J]. Decision Support Systems，2012，52（2）：406-414.
[3] 2014 ERP report. A Panorama consulting solutions research report[R]. Panorama Consulting Solutions，2014.
[4] Basu R，Upadhyay P，Das M C，et al. An approach to identify issues affecting ERP implementation in Indian SMEs[J]. Journal of Industrial Engineering and Management，2012，5（1）：133-154.
[5] Tsai W H，Chou Y W，Leu J D，et al. Investigation of the mediating effects of IT governance-value delivery on service quality and ERP performance[J]. Enterprise Information Systems，2015，9（2）：139-160.
[6] 宗威，吴锋，马超. 基于双因素理论的 ERP 实施成功认知差异研究[J]. 工业工程与管理，2013，18（5）：80-87，92.
[7] Rao D，Gudivada V N，Raghavan V V. Data quality issues in big data[C]. 2015 IEEE International Conference on Big Data. Santa Clara，2015.
[8] 李慧. 浅谈企业物料主数据标准化管理[J]. 企业改革与管理，2016，1（18）：3-4.
[9] 徐喆昊，吴共庆，胡学钢. 基于同义实体识别的 Web 信息集成[J]. 计算机系统应用，2015，24（9）：35-42.
[10] Selvi P. An analysis on removal of duplicate records using different types of data mining techniques：a survey[J]. International Journal of Computer Science and Mobile Computing，2017，6（11）：38-42.
[11] Xiao C，Wang W，Lin X，et al. Efficient similarity joins for near-duplicate detection[J]. ACM Transactions on Database Systems，2011，36（3）：1-41.
[12] 叶焕倬，吴迪. 相似重复记录清理方法研究综述[J]. 现代图书情报技术，2010，26（9）：56-66.
[13] 姜华，韩安琪，王美佳，等. 基于改进编辑距离的字符串相似度求解算法[J]. 计算机工程，2014，40（1）：222-227.
[14] Vatsalan D，Christen P，Verykios V S. A taxonomy of privacy-preserving record linkage techniques[J]. Information

Systems，2013，38（6）：946-969.
[15] 何海江. 一种适应短文本的相关测度及其应用[J]. 计算机工程，2009，35（6）：88-90，96.
[16] 华秀丽，朱巧明，李培峰. 语义分析与词频统计相结合的中文文本相似度量方法研究[J]. 计算机应用研究，2012，29（3）：833-836.
[17] Gad W K，Kamel M S. Enhancing text clustering performance using semantic similarity[C]. 11th International Conference on Enterprise Information Systems，2009：325-335.
[18] Li J. A study on noisy typing stream analysis using machine learning approach[J]. Enterprise Information Systems，2012，102：149-161.

Duplicated Material Records Identification in ERP based on Different Features Vector

ZONG Wei[1]，LIN Songtao[1]，HAO Heng[1]，WU Feng[2]
（1. School of Economics and Management，Xidian University，Xi'an 710071，China；2. School of Management，Xi'an Jiaotong University，Xi'an 710049，China）

Abstract Duplicated records are very pervasive during the initial implementing stage of an ERP system. What's worse，the variety of material names makes it difficult and complicated to identify duplicated material records. Most of the traditional methods are character-based or vector-space based which ignore the semantic similarity between compared records and couldn't identify the duplicated material records effectively. Based on the characteristics of material names，material names are firstly reversed，sorted，indexed and transformed into the vectors of different words；then a machine learning method，probabilistic neural network（PNN） is applied to learn the semantic similarity between material names so as to identify duplicated material records.

Keywords Different features vector，Duplicated material records，Semantic similarity，Data quality，ERP

作者简介

宗威（1986—），女，西安电子科技大学经济与管理学院讲师、硕士生导师，研究方向为数据质量、多源异构数据融合等。E-mail：weizong@xidian.edu.cn。

林松涛（1998—），男，西安电子科技大学经济与管理学院硕士研究生，研究方向为数据集成、真值发现等。E-mail：1462365946@qq.com。

蒿恒（1991—），男，西安电子科技大学经济与管理学院硕士研究生，研究方向为元数据集成、数据质量等。E-mail：haoheng009@163.com。

吴锋（1964—），男，西安交通大学管理学院教授、博士生导师，研究方向为智能制造管理、外包决策与控制等。E-mail：fengwu@mail.xjtu.edu.cn。