doi: 10.7690/bgzdh.2019.11.009

基于自然语言处理方法的 S 系列标准间数据映射关系

张　琦[1]，王　正[2]，朱兴动[3]，范加利[2]
（1. 海军航空大学青岛校区研究生队，山东 青岛 266000；2. 海军航空大学青岛校区六系，山东 青岛 266000；
3. 海军航空大学，山东 烟台 264000）

摘要：为解决 S 系列标准映射关系寻找难度大等问题，使用基于自然语言处理方法进行研究。分析 S 系列各标准数据元素的特点，并与人工寻找方法所得的映射关系进行对比验证。实验结果表明：该方法不仅能大幅提高寻找映射关系的效率，还能在较小的范围内寻找出所有的准确匹配对，便于人工筛选。

关键词：S 系列综合保障信息；自然语言处理；数据元素；映射关系
中图分类号：TJ07　**文献标志码：**A

Data Mapping Relationship Between S Series Standards Based on Natural Language Processing Method

Zhang Qi[1], Wang Zheng[2], Zhu Xingdong[3], Fan Jiali[2]
(1. *Brigade of Postgraduate*, *Qingdao Branch*, *Navy Aviation University*, *Qingdao* 266000, *China*;
2. *No.*6 *Department*, *Qingdao Branch*, *Navy Aviation University*, *Qingdao* 266000, *China*;
3. *Navy Aviation University*, *Yantai* 264000, *China*)

Abstract: In order to solve the problem of difficult finding in searching for the S series standard mapping relationship, research it by using natural language processing methods. Analysis of the characteristics of the standard data elements of the S series, and compared with the mapping relationship obtained by the manual search method. Experimental results show that this method can not only greatly improve the efficiency of finding mapping relationships, but also possible to find all the exact matching pairs in a small range, which is convenient for manual screening.

Keywords: S series integrated logistic support information; natural language processing; data element; mapping relations

0　引言

为了规范对武器装备的综合保障过程，不断完善欧洲 S 系列综合保障标准，先后出版了规范技术出版物编制的 S1000D[1]，规范综合保障物料供应与采购过程数据的 S2000M[2]，规范整个综合保障过程及数据的 S3000L[3]，规范预防性维修过程及数据的 S4000P[4]，规范装备使用过程产生的反馈数据 S5000F[5]，同时，规范训练内容的 S6000T 也即将出版。这些标准逐渐形成了 S 系列综合保障标准框架。它功能强大，在规范装备综合保障各个方面的同时，使综合保障工作体系化，并促使综合保障信息的集成化发展。在 S 系列综合保障框架下，标准与标准间联系紧密，存在数据共享的趋势。S 系列标准体系对于我国装备综合保障信息间交互性差、数据重用率低的现状具有借鉴意义。为了进一步确定可用于任意标准间共享的具体信息，有必要寻找出 S 系列标准间的数据映射关系。同时，S1000D 的数据元素与其他标准的数据元素命名方式不同，加大了寻找映射关系的难度。

自然语言处理方法中的 word2vec 方法与 doc2vec 方法是基于对词的向量化操作，通过对语言模型的训练，最终计算出所有文本中每个词语的词向量，根据词向量间的距离寻找意义相同的其他词语，很适合用于寻找标准间这种数据元素词语结构不同但意义相同的匹配对寻找。

1　S 系列各标准数据元素特点

1.1　S1000D

S1000D 规范的数据以可扩展标记语言（extensive markup language，XML）形式表示。XML 是一种通用语言规范，以树形结构定义数据。树节点表示每个数据元素，树枝表示相连数据元素间的关系。XML 在描述数据内容的同时能突出对结构的描述，从而体现出数据之间的关系。

如图 1，所有的元素都是该树形结构的一个节点，节点与节点之间依靠连线表示相关关系。〈itemSeqNumber〉即为这个树形结构的母节点，其他在连线另一端的节点都认为是〈itemSeq

收稿日期：2019-06-12；修回日期：2019-07-29
作者简介：张　琦（1994—），男，湖南人，硕士，从事武器装备信息化研究。E-mail: 35368504@qq.com。

Number〉的子节点，母节点与子节点的关系是包含关系，因此每个子节点都是〈itemSeqNumber〉的一部分。其中，子节点也可以是其他节点的母节点。

图 1　节点〈itemSeqNumber〉XML 树结构

通过观察发现：每个节点的名称都是几个英文常规单词拼接而成的短语，类似英文中的合成词，而且这些短语的词性基本都是名词词性，表达的是对某一种数据元素的定义。但大多数短语由于未被英语官方定义过，不能称作英文单词，其含义也不能简单地用其中任何一个单词来概括。由于这些短语的独特性，理解它们变得困难，同时找到这些短语与其他标准定义数据中相同意义的数据元素则难度更大。

1.2　其他 S 系列标准

统一建模语言(UML)是一个支持模型化和软件系统开发的图形化语言。如图 2 所示，除 S1000D 以外的 S 系列标准都采用 UML 中的这类图表示。UML 是面向对象的分析与设计方法开始盛行之后的产物，以图形化的方式表现数据模型，易于理解。以 S2000M 标准第 14a 章第 15.3 节的零部件供应数据模型为例，描述了零部件的供应信息，内容包括零部件的基本信息类(HardwarePartAsDesigned)，零部件保障数据(HardwarePartAsDesignedSupportData)，价格变动数据(PriceBreakData)等数据类。

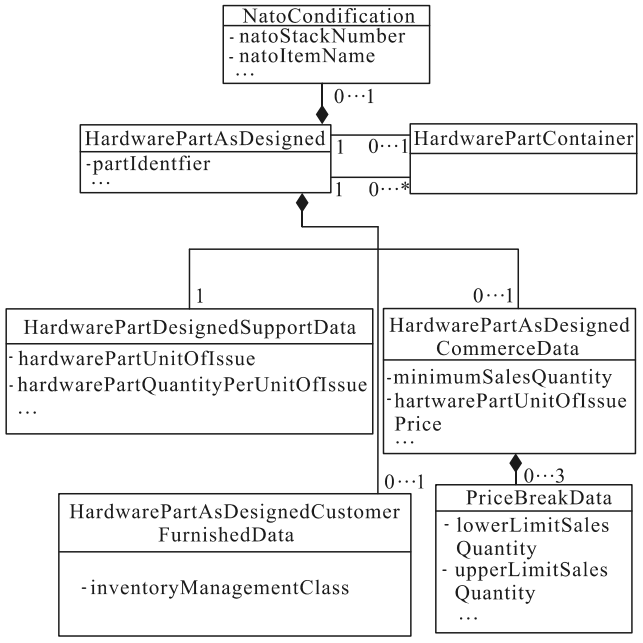

图 2　S2000M 零部件供应数据模型类图

通过观察发现，每个 UML 类图描述的数据模型都由几个类和这几个类的关系组成。类与其中属性的名称都是几个单词拼成的合成词，菱形箭头相连的 2 个类是聚合关系，如 HardwarePartAsDesignedSupportData 与 HardwarePartAsDesigned。每个类中包含了许多不同数据类型的属性。属性拥有自己的属性值。一个类可以理解成一张关系表，属性即为表头。

2　自然语言处理方法简介

2.1　word2vec 方法

无论中文还是英文词语，都是字符数据，不能直接拿来进行计算。自然语言处理领域一般将词语、句子之类的字符数据向量化处理，用向量代替词语并输入计算机进行计算。词语与词语之间的距离可以理解为向量与向量之间的距离，距离越大，词语相似度越小。为了在计算中考虑词语的语义信息，分布假说应运而生。它认为：上下文相似的词语语义也相似。寻找词语语义相似的问题，就演变成寻找某个词的上下文词语相似问题。一般词向量的产生需要训练才能得到，是训练语言模型过程中的副产物。word2vec 的训练过程可以看作是通过神经网络的机器学习算法来训练 N-gram 语言模型，并在训练过程中求出 word 所对应 vector 的方法。根据语言模型的不同，又可分为“CBOW”和“Skip-gram”2 种模型[6-7]。

2.1.1 CBOW 模型

连续词袋模型(continuous bag-of-word model，CBOW)是一个 3 层的神经网络。如图 3 所示，该模型的功能是输入已知上下文，输出对当前单词的预测。第 1 层是输入层，首先选定一个窗口大小，设其中间的单词为每一个词随机初始化一个 $\boldsymbol{K}$ 维向量，则 CBOW 模型的输入是上下文窗口内除了当前词的词向量；中间层将上下文词的向量累加(或者求均值)得到中间词的词向量；第 3 层是 Huffman 树结构。Huffman 树是一种特殊的二叉树结构，最多有 2 个有序子树(左右子树顺序不能更换)。叶节点(分枝到最底层无法继续分支的节点)代表语料里所有的词，对于一个叶节点，从根节点沿着树枝走向该叶节点，如果记左子树为 1，右子树为 0，就会产生一个唯一编码，例如“010”，代表着从根节点指向该叶节点的唯一路径(如图 3)。

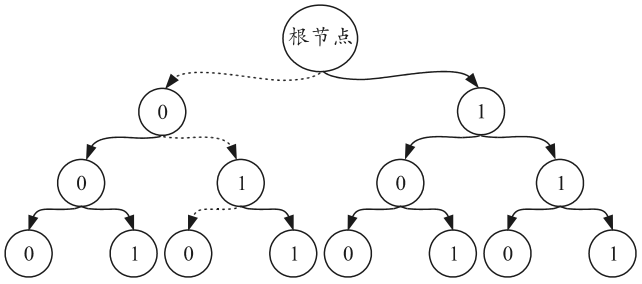

图 3 “010”的 Huffman 树结构

根据语料库建立词汇表 V。V 中的所有词均初始化一个 $\boldsymbol{K}$ 维向量，并根据词频构建 Huffman 树。将语料库中的文本依次进行训练，以一个文本为例，将单词 w_t 的上下文窗口内的词向量输入模型，由中间层处理(求和或求平均)之后得到 $\boldsymbol{K}$ 维的中间向量 $\boldsymbol{w}_{\text{new}}$。$\boldsymbol{w}_{\text{new}}$ 在 Huffman 树中沿着某个特定的路径到达某个叶子节点(即当前词 w_t)。由于已知 w_t，则根据 w_t 的 Huffman 编码，可以反向确定从根节点到叶节点的正确路径，也确定了在路径的所有非叶结点上应该作出的预测。按照此方法，从中间向量开始，沿着路径走向代表当前词的叶结点，给定如下训练单词序列 $w_{t-k},w_{t-k+1},\cdots,w_{t+k-1},w_{t+k}$，则 w_t 在当前网络下的概率为：

$$P(w_t \mid w_{t-k},\cdots,w_{t-1},w_{t+1},\cdots,w_{t+k})。$$

然后将上下文的词向量求平均或者求和作为特征(中间)向量，预测窗口中的中间位置单词。目标函数即当上下文出现词序列 $w_{t-k},\cdots,w_{t-1},w_{t+1},\cdots,w_{t+k}$ 时，中间被预测的词为 w_t 的概率。设上下文词序列为 c，则目标函数公式为

$$\sum_{(w,c)\in D}\log P(w,c)。$$

经过模型计算得到的解不符合概率值范围要求。为了得到 $P(w,c)$，使用 softmax 函数将其归一化，使值控制在区间 $[0,1]$，归一化后的目标函数计算公式如下：

$$\sum_{(w,c)\in D}\log\left[\frac{\exp(e'(w)^{\mathrm{T}}x)}{\sum_{w'\in V}\exp(e'(w')^{\mathrm{T}}x)}\right]。$$

得到目标函数后，可以采用梯度下降优化算法调整路径中非叶结点的参数，以及最终上下文词的向量，使得实际路径向正确路径靠拢，经过 n 次迭代收敛后，即可得到每个词的最终向量表示。CBOW 模型训练过程如图 4 所示。

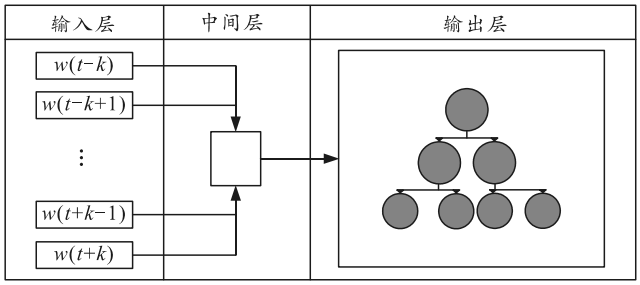

图 4 CBOW 模型训练过程

2.1.2 Skip-gram 模型

与 CBOW 模型的思路正好相反，Skip-gram 模型由限定当前词来预测上下文的每一个词。同样，Skip-gram 模型也是一个 3 层神经网络，也同样分为输入层、中间层和 Huffman 树层。与 CBOW 模型不同的是，Skip-gram 模型输入层不再是多个词向量，而是只有一个词向量，所以不用做中间层的向量处理。具体的训练过程如图 5 所示。

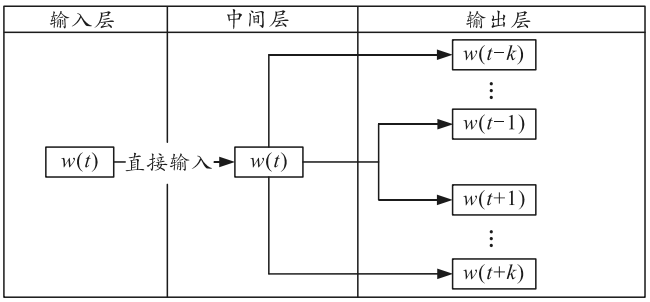

图 5 Skip-gram 模型训练过程

无论是 CBOW 还是 Skip-gram 模型，其实都是分类模型。2 种模型都将上下文词语的关系作为建模，得出词向量的重要考量，即考虑了词之间的语义信息，加上大量训练文本作为建模输入，使得迭代的效果能够达到需求，实际路径与正确路径充分

拟合。这样得到的庞大词向量集合就可通过计算向量的各种距离，来得到词与词间的距离和相平似度[8-13]。

2.2　doc2vec 方法

word2vec 方法表示的词向量不仅考虑了词之间的语义信息，而且压缩了维度。如果只比较词与词之间的相似度，方法的使用面就比较窄，大多数情况下需要了解句子与句子，甚至文本与文本之间的相似度，据此便可进一步实现文档分类、情感分析等智能化的功能[14-16]；因此，必须求得句子或者文本段落的向量表示，虽然可以直接将句子或者文本段落中所有词向量取均值作为其向量表示，但是会忽略词与词之间的排列顺序对句子或文本信息的影响。例如，“A 送 B 一个苹果”与“B 送 A 一个苹果”这 2 句意思正好相反，用 word2vec 方法就认为这 2 句话相同。在 word2vec 模型的基础上进行改进，得到 doc2vec 模型[17]。与 word2vec 一样，doc2vec 也有 2 种模型，分别为 Distributed Memory(DM)和 Distributed Bag of Words(DBOW)。DM 模型在给定上下文和文档向量的情况下预测单词的概率，DBOW 模型在给定文档向量的情况下预测文档中一组随机单词的概率。设一句话中 3 个词的词向量为 w_1、w_2 和 w_3，则预测下一个词的词向量 DM 模型结构如图 6 所示。

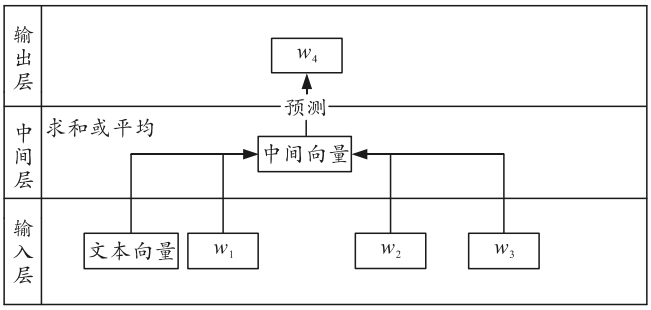

图 6　DM 模型训练过程

与 CBOW 模型不同的是，DM 模型在训练时增加了一个与其他词向量长度相同的文本向量。这里引入文本 ID 的概念，每个句子或段落都被唯一标识一个文本 ID，首先将每个文本 ID 和语料库中的所有词都初始化一个 $\boldsymbol{K}$ 维的向量，然后将文本向量和上下文词的向量输入模型，中间层将这些向量处理(相加或取均值等)得到中间向量，作为输出层的输入。在一个文档的训练过程中，文本 ID 保持不变，共享着同一个文本向量，相当于预测单词概率利用了整个文本的语义。在预测单词的过程中，给将要预测的文档重新分配一个文本 ID，其他参数保持不变，重新使用随机梯度下降算法训练预测的文档。待误差收敛后，即得到预测文档的文本向量。

DBOM 模型结构如图 7 所示。

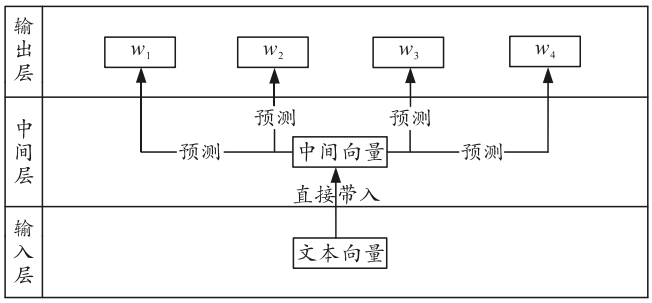

图 7　DBOW 模型训练过程

类似于 CBOW 与 Skip-gram 模型，DM 模型与 DBOW 模型的区别在于，DBOW 模型的输入只有文本向量，预测的是该文档中随机抽样词的概率分布。

3　寻找标准间映射关系实验及效果对比

语料是整个实验训练的必备材料与最根本要素。实验选取维基百科英文语料作为基础，S 系列综合保障标准原文作为增量训练输入。如图 8 所示，维基百科是一个网络百科全书项目，特点是内容自由、编辑自由，是目前全球网络上最大且最受欢迎的百科工具。

图 8　中文维基百科官网

3.1 word2vec 效果及局限性

文本预处理(格式转换、去除停用词、文本纠错)之后的语料可以直接用于 word2vec 文本训练，训练过程中使用到的 word2Vec 函数参数设置如表 1。

表 1 Word2Vec 参数设置及含义

参数	含义
size=300	表示设置训练生成的词向量维数为 300
window=10	表示一个窗口包含 10 个词
min_count=5	表示用于训练生成词向量的文本出现次数必须不小于 5 次
workers=multiprocessing.cpu_count	表示用 cpu 所能达到的最多进程来训练词向量
sg=0	表示训练使用的模型是 CBOW
iter=5	迭代次数设置为 5 次

设置好参数后便可以开始训练，通过提取文本，将词定量打包输入模型进行训练之后，得到一个体型庞大的 model 格式文件。准备 2 个空文本文件，一个用来逐行存放 S1000D 图解零部件目录数据模块的所有数据元素名称，一个用来存放 S2000M 与 SX002D 所有数据元素的交集(S1000D 元素节选如图 9)，检索 2 个文本逐行的所有匹配对，并输入 model 文件计算每个匹配对的语义相似度。

authorityExceptions
authorityInfo
authorityInfoAndTp
authorityNotes
behavior
brexDmRef
buyerFurnishedEquipFlag
calibrationMarker
catalogSeqNumber

图 9 S1000D 数据元素节选

通过统计，相似度大于 0.5 的只有 1 对，大于 0.4 的有 50 对，大于 0.3 的有 304 对，不仅没有出现高相似度的匹配对，而且准确率经过粗略统计也很低，无法根据结果寻找符合要求的匹配对。究其原因，方法和实验过程完全按照标准进行；因此，只能从语料上寻找原因。语料来源为维基百科和 S 系列综合保障标准英文原文。维基百科被许多实验证实了有效性，那么问题来源就可以定位到标准原文上。标准原文的数据体量远小于维基百科的体量。在很小的范围内，标准中描述数据模型中数据元素的内容只占该范围的一部分，同时每个数据元素的名称都不是一个意义简单的单词，而是一个合成词，可以看作一个短语，承载的语义大于一个真正意义上的单词。原因总结为，较少的相关语料和复杂的词语含义导致了结论的不准确。

3.2 doc2vec 方法效果及分析

doc2vec 方法不仅能求词向量，还能进一步求组成的词组、句子甚至段落的向量(暂且叫作文本向量)，正好能够解决 S 系列标准数据模型中数据元素名称含义复杂的问题。同样准备 2 个文本文件，一个按行存放从 S1000D 数据元素描述内容剥离出的数据元素定义语句，一个按行存放 S2000M 数据元素定义语句(语句节选如图 10)，再次以同样的语料和参数设置来训练适用 doc2vec 的模型(其中设置参数 sg=0，表示训练使用的模型是 DM 模型)。

authority exceptions
authority information
authority block and technical publications information
authority notes
link behavior information
business rules reference
buyer furnished equipment flag
calibration marker
catalog sequence number

图 10 以 S1000D 为例的数据元素定义语句节选

训练之后得到以 d2v 为后缀的模型文件和若干个以 npy 为后缀的文件(如图 11)。它们都是体量大的文件。加载模型文件并检索 2 个文本逐行的所有匹配对，通过 doc2vec 内置的用于计算短文本相似度的 n_similarity (w_{s1} , w_{s2})函数就能够计算每个匹配对的语义相似度，其中 w_{s1} 和 w_{s2} 分别为 2 个字符串型列表，分别包含了 2 个标准的数据元素定义语句，并以分词的形式表示。如“authorityInfo”的定义语句“authority information”表示为[“authority”,”information”]。

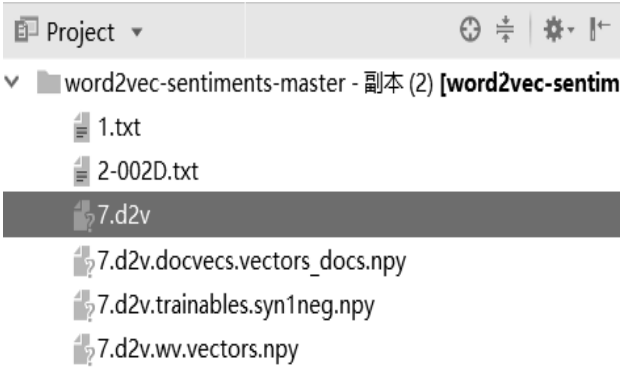

图 11 doc2vec 方法训练所得文件

通过将定义语句带入模型，从相似度值为 0.050 开始，以 0.100 为相似度分割区间，计算各匹配对的相似度。计算结果保留 3 位有效数字，并筛选出各分数段准确匹配对的个数及遗漏个数，制成曲线如图 12。经过与人工筛选出的所有准确匹配对对

比，得到匹配对 genericPartData 与 partName 相似度值为 0.450，是所有准确匹配对相似度值中的最小值，而匹配对 hazardousClass 与 hardware-PartHazardousClass 相似度值 0.890 为最大值。通过条件筛选可知：大于 0.450 的匹配对个数为 429 对，相比于人工筛选方法需要筛选 10 000 多对匹配对的情况，减少了近 96%的工作量。各小区间内的匹配对个数及准确匹配对个数如下：

[0.350,0.450)：该区间范围内共有 900 对匹配对，其中准确的匹配对为 0 对；

[0.450,0.550)：该区间范围内共有 314 对匹配对，其中准确的匹配对为 3 对；

[0.550,0.650)：该区间范围内共有 88 对匹配对，其中准确的匹配对为 4 对；

[0.650,0.750)：该区间范围内共有 17 对匹配对，其中准确的匹配对为 4 对；

[0.750,0.850)：该区间范围内共有 9 对匹配对，其中准确的匹配对为 4 对；

[0.850,0.950)：该区间范围内共有 1 对匹配对，其中准确的匹配对为 1 对。

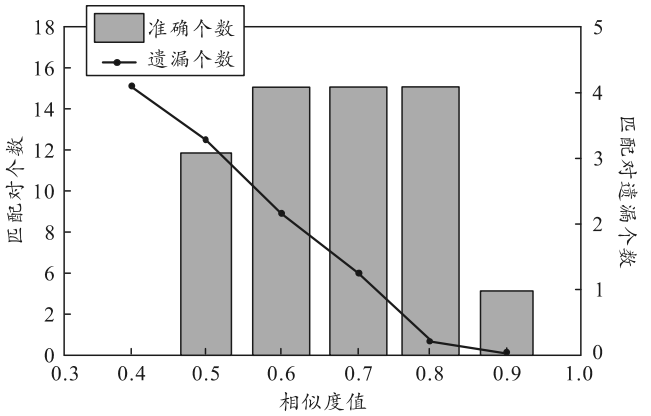

图 12　doc2vec 方法各相似度区间的准确个数及遗漏个数走势

可见，使用 doc2vec 方法能够解决 word2vec 难以生成短文本、句子甚至段落向量，进而解决短文本、句子和段落的相似度等一系列问题，结果比不包含语义的字符串匹配在准确度上有了大幅提升，方便进一步的人工筛选；因此，该方法寻找 S1000D 与其他 S 系列标准间相同意义的数据元素是可行的。S1000D 图解零部件数据元素与 S2000M 数据元素的映射关系如表 2 所示。

表 2　S1000D 图解零部件数据元素与 S2000M 数据元素的映射关系表

S1000D 数据元素	S2000M 数据元素	对应 S2000M 数据元素属性	相似度(0～1)
authorityInfo	ChangeAuthorization		0.651
ChangeAuthority			0.792
Hazardous Class		hardwarePartHazardousClass	0.890
language		MessageLanguage	0.719
genericPartData	PartAsDesigned		0.450
shortName			0.639
name		partName	0.845
partKeyword			0.764
techName			0.564
OverlengthPartNumber			0.556
partNumber		partIdentifier	0.677
limitedPartNumber			0.572
descrForPart	hardwarePartAsDesignedDesignData	hardwarePartSize hardwarePartWeight ⋮	0.480
Special Storage		hardwarePartSpecialStorageRequirement	0.767
modelVersion		productVariantIdentifier	0.471
initialProvisioningProject		hardwarePartProvisioningCategory	0.701

4　结束语

S1000D 与其他 S 系列标准的数据元素在命名方式与数据格式上都不相同。为了能够将其他综合保障信息系统的数据吸收到 S1000D 规范的 IETM 中，需要找到它们之间的映射关系。笔者主要研究了寻找 S1000D 与其他 S 系列标准相同意义数据元素的方法。通过自然语言处理方法，计算标准间数据元素的语义相似度并得到映射关系。实验所用的 word2vec 算法是通过训练维基百科英文语料与 S 系列标准原文，得到一个将词语全部向量化的模型。在基于这个模型计算数据元素相似度时，由于 S 系列标准数据元素含义复杂的限制，实验效果并不理想；于是取各数据元素的定义语句，采用支持将文本向量化的 doc2vec 方法，训练出新的模型，将数据元素定义语句向量化。通过计算代表各定义语句的向量距离得到数据元素相似度，效果较好，人工筛选工作量也在可接受范围内。

（下转第 50 页）

4 结论

笔者设计的平台能提升天平的自动化水平。该平台实现了基于 Web 服务的天平设计模式，从根本上提高了软件使用的方便性，能充分利用银河大型服务器的强大计算能力，划分出比以前更细更高的网格，在几百上千个方案中寻求最优方案，最终提升天平设计质量和效率。

参考文献：

[1] 张海天. 杆式应变天平优化设计方法研究与应用[D]. 南京: 南京航空航天大学, 2012.

[2] LINDELL M. Finite Element Analysis of a NASA National Transonic Facility Wind Tunnel Balance[A]. NASA/CP-1999-209101/ PT2, 1999: 595-606.

[3] 解亚军, 叶正寅, 杨中艳. 盒式应变天平优化设计与有限元分析[M]. 北京: 机械科学与技术出版社, 2011(12): 1974-1976.

[4] 向光伟, 王杰, 史玉杰, 等. 基于 iSIGHT 优化风洞应变天平设计方法研究[J]. 实验流体力学, 2015, 29(5): 45-49, 59.

[5] 王博文, 秦建华, 黄叙辉, 等. 风洞流场控制系统规范化研究与应用[J]. 兵工自动化, 2018, 37(6): 33-37.

[6] 中国人民解放军总装备部. GJB2244A—2011 风洞应变天平规范[S]. 北京: 总装备部军标出版发行部, 2011.

[7] 吴坤安, 严宣辉, 陈振兴. 一种基于 Pareto 排序的混合多目标进化算法[J]. 计算机工程与应用, 2015, 51(1): 62-68.

[8] 谢涛. 多目标优化与决策问题的演化算法[J]. 中国工程科学, 2002, 13(2): 80-91.

[9] 田正波, 杨家军, 史玉杰. 一种新的风洞试验支撑机构横向弹性角校准方法[J]. 兵器装备工程学报, 2017(8):32-35.

****

（上接第 43 页）

参考文献：

[1] ASD/AIA S1000D. International specification for technical publications utilizing a common source database[Z]. Isue4.1. America: ASD-AIA, 2009.

[2] ASD/AIA S2000M. International specification for material management[Z]. Issue6.0. America: ASD-AIA, 2015.

[3] ASD/AIA S3000L. International procedure specification for logistics support analysis[Z]. Issue1.1. America: ASD-AIA, 2014.

[4] ASD/AIA S4000P. International specification for developing and continuously improving preventive maintenance[Z]. America: Issue1.0. ASD-AIA, 2014.

[5] ASD/AIA S5000F. International specification for in-service data feedback[M]. Issue1.0. America: ASD-AIA, 2016.

[6] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient Estimation of Word Representations in Vector Space[Z]. Computer Science, 2013.

[7] MIKOLOV T, SUTSKEVER I, CHEN K, et al. Distributed representations of words and phrases and their compositionality[C]//International Conference on Neural Information Pressing Systems. America: Curran Associates Inc., 2013: 3111-3119.

[8] 任郭珉. 基于文本挖掘的药用植物数据库的建立及网络药理学分析[D]. 北京: 北京协和医学院中国医学科学院; 北京协和医学院; 清华大学医学部; 中国医学科学院, 2014.

[11] 王文琪, 胡炎, 赵娜, 等. 基于距离权重向量优化模型的虚端子自动连接方法[J]. 电网技术, 2018, 42(1): 346-352.

[12] 王逍翔, 杜庆治, 赵继东, 等. 基于词信息量加权的地理 POI 数据融合新方法研究[J]. 软件导刊, 2018(3): 41-44, 49.

[13] 李光军. 女子赛艇训练数据挖掘研究[D]. 武汉: 武汉大学, 2015.

[14] 任高山. 基于微信公众平台的文本情感分析研究[D]. 南昌: 南昌航空大学, 2018.

[15] 王峰, 林丽珊, 刘毅. 基于群组平台知识圈的精准信息推荐[J]. 现代情报, 2018(7): 74-80.

[16] 于凤, 郑雨晴, 郑德权, 等. 面向选择类题型求解的相似问题发现研究[J]. 计算机工程与应用, 2018(15): 120-125.

[17] LE Q V, MIKOLOV T. Distributed Representations of Sentences and Documents[J]. Computer Science, 2014, 4: II-1188.