基于相似度的装备数据聚合方法

杨　杉
(32683 部队,沈阳,110000)

摘要　现代战争需要对多源异构的装备数据进行高效集成。针对不同来源数据中装备名称不一致的问题,设计了装备数据的聚合模型和聚合流程,在综合分析现有算法的基础上,结合装备名称特点为该模型提供了一种新的相似度匹配算法,算法将 Jaro-Winkler 和最长公共子序列相结合,以提高匹配的精度。最后通过实验进行了验证,结果表明该算法与传统相似度算法相比具有较高的适配性和鲁棒性,可以为装备数据聚合工作提供有效支撑。

关键词　装备数据;数据聚合;文本相似度;Jaro-Winkler;最长公共子序列
DOI　10.3969/j.issn.2097-1915.2023.02.013
中图分类号　TP391　　**文献标志码**　A　　**文章编号**　2097-1915(2023)02-0098-06

Research on Aggregation Method for Equipment Data Based on Similarity

YANG Shan
(Unit 32683, Shenyang 110000, China)

Abstract　Modern warfare requires efficient integration of multi-source heterogeneous equipment data. To solve the problem of inconsistent equipment names of data from different sources, the aggregation model and aggregation process of equipment data are studied and designed. Based on the comprehensive analysis of existing algorithms and the characteristics of equipment data, a new similarity algorithm is provided for the model, the algorithm combines Jaro-Winkler and the longest common subsequence to improve the matching accuracy. Finally, experiments show that the algorithm has high adaptability and robustness compared with the traditional similarity algorithm, and can provide effective support for equipment data aggregation.

Key words　equipment data; data aggregation; text similarity; Jaro-Winkler; longest common subsequence

信息融合和信息主导已成为现代战争体系作战能力生成与发挥的重要机理[1],随着装备信息化建设的不断深入,如何将多源异构的装备数据进行有效的融合集成已经成为不可回避的问题[2]。装备数据聚合的主要矛盾是异源数据装备命名不唯一,该问题导致数据汇聚后难以直接使用,后期整编、维护、管理成本巨大,亟需寻找一种行之有效的装备名称匹配聚合方法。当前数据聚类、聚合的算法模型众多,在不同应用中效果迥异[3],但在装备数据聚合问题上尚无典型的研究案例和行之有效的实施方法。因此,本文针对装备数据聚合进行研究,根据装备名称的特点设计了装备数据的聚合模型和聚合流

收稿日期:2022-10-08
基金项目:国家自然科学基金(71701205)
作者简介:杨　杉(1985—),男,辽宁新民人,高级工程师,研究方向为数据理论。E-mail:ysforwin@163.com

引用格式:杨杉.基于相似度的装备数据聚合方法[J].空军工程大学学报,2023,24(2):98-103. YANG Shan. Research on Aggregation Method for Equipment Data Based on Similarit[J]. Journal of Air Force Engineering University, 2023, 24(2): 98-103.

程,并尝试为该模型提供一个适用性较强的相似度计算方法。

1　相关背景及概念

1.1　装备数据问题分析

“装备数据”没有明确的定义,其内涵和外延存在较多解读,有学者将“装备数据”定义为“用于描述装备自身特性和状态的数据以及装备全系统全寿命管理活动所涉及的数据的统称”[4]。装备数据涵盖范围宽泛,包含装备属性数据、装备标准数据、装备衍生数据、装备数质量数据等。作为装备数据的主体,这些数据都含有大量的装备名称,由于装备迭代更新加快、信息系统不相兼容、采集任务标准不一、人工采报内容格式不规范等原因,当前装备数据中存在大量装备名称不一致的情况,这为装备数据融合带来了困难,使装备数据难以形成统一的数据资源。

装备名称不一致是指同一装备因数据来源不同存在多种命名方式的现象,该现象会严重影响数据质量,导致融合后数据失准,造成数据分析偏差,甚至影响指挥决策。例如装备 A 可能有 A_1、A_2、A_3 等多个名称,如果直接归集使用,A 的统计数量会有缺漏,A_1 等非标准名称也将无法完成有效的信息关联,引发后期数据应用结果的混乱。所以在数据融合过程中必须识别出相同型号的装备并将其聚合。装备名称不一致问题产生原因多、发生频率高,情况也较为复杂,当前主要通过人工整编的方式进行解决,但因数据体量巨大,匹配工作效率低、容易出错。

1.2　数据聚合及模型选择

数据聚合的本质是数据匹配和对齐,是判断隶属 2 个不同数据集合中的 2 个实体是否属于同一个实体的过程[5]。对装备数据聚合而言,其核心是对装备名称文本数据的聚合,目标是通过匹配对齐,消除数据的不一致性,提升数据的稳定性和准确性。为了有针对性地解决装备数据的聚合问题,本文对当前的主流方法进行了研究分析。

数据聚合研究因数据时代的到来而备受关注[6-8],当前数据聚合的主流方法可分为 3 类:

1)基于概率的统计聚合模型。该模型是通过概率和权重的计算实现数据与主题词的聚合;

2)监督聚合模型。此模型是依托构建的人工神经网络模型实现数据聚合;

3)无监督聚合模型。该模型针对数据本身进行研究,主要通过文本相似度计算实现信息的聚类。

前 2 类模型针对的是长文本的数据聚类,主要应用于网络信息的检索,重点是对数据的数量和特征进行聚合。装备数据聚合因装备名称较短,难以统计词频和提取特征,统计模型和机器学习模型并不适用。故本文首先锁定无监督聚合模型作为解决方向。

无监督聚合模型又分为“计算字符距离”和“计算语义相似度”2 类。前者是将文本数据作为字符序列进行量化分析,利用一些距离来衡量二者的相似度;后者一般需要依赖外部语义词典进行知识匹配来完成相似度计算[9-10]。装备的命名主要由阿拉伯数字、英文字母、汉字及少量罗马字母组成,构成复杂、不易分词且缺少专用的外部词典;同时,装备名称特征具有稀疏性,容易对语义相似度计算的准确性造成较大的负面影响,所以不建议使用语义相似度进行聚合。综上所述,本文在无监督聚合模型的基础上,进一步确定采用基于字符距离计算相似度的方法解决装备数据聚合问题。

2　装备数据聚合模型及流程

2.1　装备数据聚合模型

装备数据聚合是指依据一定的规则将同一个装备对象的多个名称进行匹配对齐,根据匹配结果对待聚合数据集进行整合的过程。假定装备数据集 DB 中的 2 条数据 D_1 和 D_2,则 D_1 和 D_2 匹配可定义为:

$$Align(D_1,D_2)=\begin{cases}1,\text{if } Sim(n_1,n_2)>\delta\\0,\text{otherwise}\end{cases},n_1\in D_1,n_2\in D_2 \tag{1}$$

式中:n 为装备实体名称;Sim 表示 2 条数据中装备名称的聚合相似程度,Sim 值越大,n_1 和 n_2 越相似。可根据聚合要求设定阈值 δ,如果 $Sim>\delta$,则对 2 条数据进行聚合处理。

装备名称具有规范性,在聚合过程中应向标准名称看齐,规范的装备名称来源于装备字典,所以如果聚合任务明确了装备字典 DD,可先将 DB 中的具体装备名称 n_i 与 DD 进行匹配,选取 DD 中与 n_i 相似度最高的名称作为匹配成功项,记为 ADD_i,即 $Sim(D_i,ADD_i)=Sim_{\text{Max}}$。因该匹配结果仅取最大值,本文称其为唯一性匹配。此时聚合的匹配定义由阈值判断转化为唯一性判别:

$$Align(D_1,D_2)=\begin{cases}1,\text{if } ADD_1=ADD_2\\0,\text{otherwise}\end{cases},n_1\in D_1,n_2\in D_2 \tag{2}$$

2.2 装备数据聚合流程

装备数据聚合流程因装备数据种类不同而略有差异，但核心思想一致，本文将装备数据聚合流程分为 3 个阶段：

Step 1 装备名称的提取和预处理。将装备名称从所有需要聚合装备数据集中全部提取出来，形成文本数据集，再对装备名称文本集进行剔重和清洗等预处理，以提升聚合的质量和效率；

Step 2 装备名称的匹配对齐。采用合适的相似度算法对装备的名称进行匹配计算，如果有明确装备字典，应将装备数据中的名称严格与装备字典进行唯一性匹配，否则通过设定相似度阈值 δ 完成匹配对齐；

Step 3 装备数据整合。装备名称匹配对齐后，首先应按标准统一装备名称，使装备名称具备唯一性，如果在聚合任务中没有明确的装备字典，建议选取信息较为完整的装备名称作为聚类后统一的标准。然后根据装备数据结构确定组合主键（一般包含装备名称，隶属部门，装备状态等），对数据进行删除重复项、整合同类数据等操作，进而得到最终聚合后的数据集合。

整个过程如图 1 所示。

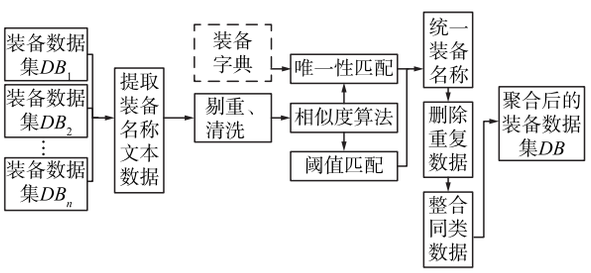

图 1 装备数据的聚合流程

分析装备数据聚合过程可知，其核心任务是对装备名称的匹配对齐，所以本文研究的重点是匹配所需的相似度算法。

3 装备数据聚合相似度算法设计

基于字符距离的相似度算法有较为丰富的研究成果，当前的主流算法有欧氏距离、余弦距离[11]、最小编辑距离[12]、Jaro-Winkler 距离[13]等。这些相似度算法各有特点，没有明显优劣。其中，最小编辑距离算法、余弦相似度等算法比较适合以分词为单位的计算，Jaro-Winkler 算法更适合以字符为单位的计算[14]。装备名称当前缺少专用“词袋”难以分词，故 Jaro-Winkler 算法相对更适用于装备名称的相似度计算，但其加强对相同前缀评估的权重不符合装备名称的特点需求，因此本文尝试在该算法的基础上进行改进。

装备型号命名一般用字母、数字和文字的组合表示该装备的基本性能、规格和产品种类。分析装备数据名称不一致现象，多数是由品牌、年代、型号、计量单位的不同写法造成的。虽然这些具体标识不完全一致，但内容本质上是相同的，所构成的装备名称文本也分散存在大量共有子域，例如“1954 年式 7.62 毫米手枪”“54 式手枪”和“54 式 7.62 mm 手枪”就拥有“54”“7.62”和“手枪”等相同子字符串。据此，本文在 Jaro-Winkler 的基础上增加对最长公共子序列的度量，具体算法如下：

首先引入配窗口概念，对于字符串 s_1 和 s_2，二者之间的匹配窗口记为 MW(matching window)：

$$MW=\left\lfloor\frac{\text{Max}(|s_1|,|s_2|)}{2}\right\rfloor-1 \tag{3}$$

式中：$|s_1|$ 和 $|s_2|$ 表示 2 字符串的长度，当 2 个相同字符的距离不大于 MW 时认定该 2 个字符是匹配的，因此通过 MW 可过滤掉距离相对较远的相同字符。过滤后得到新的字符串 sn_1 和 sn_2，因匹配字符个数相同，将新字符串长度记为 m：

$$m=|sn_1|=|sn_2| \tag{4}$$

此时 sn_1 和 sn_2 具有相同的字符集合，但部分字符位置不同。通过交换字符的方式将 sn_1 转变为 sn_2，将交换次数记为 t_s，则换位数目是发生换位次数的一半，记为 t：

$$t=\frac{t_s}{2} \tag{5}$$

通过 m 和 t 可以计算得到 Jaro 距离 d_j：

$$d_j=\begin{cases}0 & ,m=0\\\frac{1}{3}\left(\frac{m}{|s_1|}+\frac{m}{|s_2|}+\frac{m-t}{m}\right) & ,m>0\end{cases} \tag{6}$$

Winkler 在 Jaro 的基础上，对算法进行了补充，增加了对 2 个字符串前缀相同部分的考量，相似度算法改进如下：

$$d_w=d_j+(lp(1-d_j)) \tag{7}$$

式中：p 为常量，范围为(0～0.25)，根据实际情况进行调整，默认值通常设置为 0.1；l 为 2 个字符串的起始部分存在的相同字符个数，即相同最大前缀

长度。

定义：对 2 个字符串 A 和 B 在不改变字符顺序的基础上进行字符删减操作，将得到的长度最长的相同字符序列定义为 A 和 B 的最长公共子序列 LCS(A，B)，序列长度(字符个数)为 L_{LCS}，定义 A 和 B 公共子序列相似度系数为：

$$P_{\mathrm{LCS}}=\frac{2L_{\mathrm{LCS}}}{|s_A|+|s_B|} \tag{8}$$

由 $L_{\mathrm{LCS}}\leqslant \mathrm{Min}(|S_A|,|S_B|)$ 可知 $P_{\mathrm{LCS}}\leqslant 1$，所以可以用 P_{LCS} 代替式(7)中的 l 和 p 2 个变量，将 Jaro-Winkler 算法改进得到新的 d_{AB}：

$$d_{AB}=d_j+P_{\mathrm{LCS}}(1-d_j) \tag{9}$$

考虑到 2 个系数叠加后相似度数值较大，比值结果差距较小，为了在应用中更容易设定相似度阈值，对结果进行差异放大处理，得到：

$$Sim(A,B)=2(d_j+P_{\mathrm{LCS}}(1-d_j))-1 \tag{10}$$

如果 $Sim(A,B)<0$，则认定 A 和 B 不具备匹配条件；如果 $Sim(A,B)=1$，则认定 A 和 B 相同。综合公式(6)和(10)，改进后 A 和 B 的相似度计算公式为：

$$Sim(A,B)=\begin{cases}0 & ,m=0\\ \dfrac{2}{3}\left(\dfrac{m}{|s_A|}+\dfrac{m}{|s_B|}+\dfrac{m-t}{m}\right)\dfrac{|s_A|+|s_B|-2L_{\mathrm{LCS}}}{(|s_A|+|s_B|)}+\dfrac{4L_{\mathrm{LCS}}}{|s_A|+|s_B|}-1 & ,m>0\end{cases} \tag{11}$$

改进后的算法与 Jaro-Winkle 算法相比，在规避前缀错误度量的同时，增强了对共有部分的评估的权重，更符合装备名称特性，理论上可以提高聚合匹配的准确度。

4　实验结果与分析

为了验证本文设计的算法在装备名称聚合上的匹配效果，实验分别在典型数据集和较大规模数据集中应用了该算法，并与常用的字符距离相似度算法进行了比对分析。

4.1　典型数据集实验分析

在单样本比对实验中，本文尽量选取有代表性的典型情况进行计算分析。数据集以字典中“1954 年式 7.62 毫米手枪”为基础进行扩展获取，在装备数据聚合过程中常见的装备名称不一致的情况有以下 3 种：

1)年代、口径、型号等标识简写。扩展数据为“①54 年式 7.62 毫米手枪”和“②54 年式 7.62 手枪”；

2)计量单位标准不同。扩展数据为“③1954 年式 7.62 mm 手枪”；

3)缺少部分前后缀标识。扩展数据为“④1954 年式手枪”和“⑤7.62 毫米手枪”。严格意义上缺少前后缀标识的装备名称很有可能产生歧义，例如“7.62 毫米手枪”可能存在多个年代型号与之匹配，在数据量较大时，为了提升数据质量应修正补全这些数据，在典型数据集中因样本数量较少，暂不考虑名称歧义。

以上同一装备的不同名称，统称为该装备的正例。在装备名称匹配过程中，除了正例，装备数据往往更多的是非同一装备的名称反例，为了实现装备名称的自动匹配，算法应能够区分正例和反例，为了全面分析算法，在典型数据集中添加 2 个与正例相似的反例：“⑥1959 年式 9 毫米手枪”和“⑦1954 年式 12.7 毫米穿甲燃烧弹”，用于测试算法的反向匹配。现将 7 条数据分别与装备字典中的对应的标准名称进行比对，应用本文算法结果如图 2 所示。

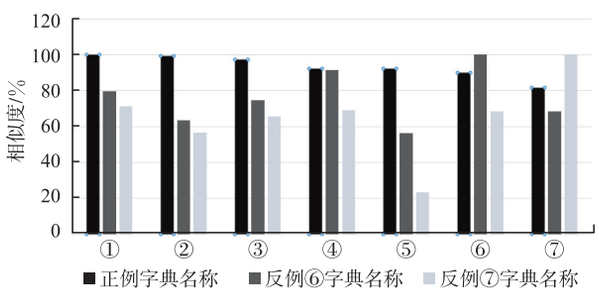

图 2　不同名称相似度对比结果

从 2 个维度对图 2 的实验结果进行分析。维度 1 是匹配分析，将 7 个数据样本分为 3 组，正例①～⑤为一组，该组数据与正例字典名称的相似度明显高于 2 个反例字典名称，采用唯一性匹配原则，正例可以匹配到正例字典名称；同样⑥和⑦各为一组，与自身对应的字典名称相似度最高，也能够完成正确匹配。维度 2 是正反例区分情况分析，比较 7 个样本与正例字典名称相似度计算结果，正例均高于反例，但反例的相似度与正例⑤和④较为接近，反观 7 个样本与 2 个反例字典名称，则明显能够进行正反例区分。综合分析结果，在典型情况的数据样本中，本文提出的算法可以实现将不一致的装备名称匹配到装备字典中对应的标准名称，也能够一定程度区分正反例，但遇到过于简单的正例和极易混淆的反例(如⑤和⑥)，区分界限不明显。

为了横向对比本文算法与各主流算法，现采用

不同算法对典型情况数据集进行计算，实验结果如表 1 所示。

表 1 6 种相似度算法对比结果

装备名称	对比名称	欧氏距离	余弦距离	最小编辑距离	Jaro 相似度	Jaro-Winkle 相似度	本文算法
1954 年式 7.62 毫米手枪	① 54 年式 7.62 毫米手枪	0.414 2	0.925 8	0.857 1	0.952 4	0.952 4	0.992 7
	② 54 年式 7.62 手枪	0.333 3	0.845 2	0.714 3	0.904 8	0.904 8	0.968 3
	③ 1954 年式 7.62 mm 手枪	0.289 9	0.801 8	0.857 1	0.904 8	1.000 0	0.972 8
	④ 1954 年式手枪	0.289 9	0.755 9	0.571 4	0.857 1	0.942 9	0.922 1
	⑤ 7.62 毫米手枪	0.289 9	0.755 9	0.571 4	0.857 1	0.857 1	0.922 1
	⑥ 1959 年式 9 毫米手枪	0.250 0	0.713 0	0.642 9	0.820 3	0.874 2	0.899 4
	⑦ 1954 年式 12.7 毫米穿甲燃烧弹	0.250 0	0.735 8	0.470 6	0.780 6	0.912 2	0.816 0

对表 1 进行分析，因为欧氏距离和余弦距离主要考虑字符出现的频率，所以这两个算法对缺少标识的情况较为敏感，正例匹配的相似度范围大，无装备字典聚合时相似度阈值设定较为困难，而且二者不考虑字符的位序，简化的正例和相近的反例相似度接近，容易错误匹配；最小编辑距离算法对字符串的长度和位序都较为敏感，正向匹配较为困难，甚至出现反例⑥的相似度高于正例⑤的相似度，也不符合装备名称聚合要求；Jaro 相似度相比前 3 个算法较稳定，但对连续相同字符不敏感导致正例相似度区间偏大，Jaro-Winkle 过于看重相同前缀导致⑦高于⑤，也不适用与装备名称的聚合；综合比较，在典型数据集中本文算法较其他算法更为稳定，匹配准确率相对较高。

4.2 较大规模数据集算法对比分析

为了验证本文算法的匹配准确率和鲁棒性，将算法应用于较大规模的装备数据集，数据集约 30 000 条装备数据，提取装备名称实例 1 221 个。为了对算法的有效性进行分析，将匹配结果进行二元分类，建立混淆矩阵：

True Positive(真正，TP)：同一装备且名称符合匹配；

True Negative(真负，TN)：不是同一装备且不符合匹配；

False Positive(假正，FP)：不是同一装备但符合匹配；

False Negative(假负，FN)：是同一装备但不符合匹配。

通过混淆矩阵可以得到算法的准确率 ACC：

$$ACC=\frac{TP+TN}{TP+TN+FP+FN} \tag{12}$$

算法的精确率 P 和召回率 R 分别为：

$$P=\frac{TP}{TP+FP},R=\frac{TP}{TP+FN} \tag{13}$$

其中精确率和召回率是 2 个不同角度对算法的分析，理想情况下，P 和 R 同时越高越好，但现实情况中，准确率与召回率往往呈反比，根据装备聚合的要求应保证尽量不错漏正例，实验的阈值确定原则为在保证召回率的前提下精确率最大化。同时为了全面分析算法的有效性，实验中引用更为综合的评价指标 F-Measure[15]，其定义为：

$$F=\frac{(\beta^2+1)PR}{\beta^2P+R} \tag{14}$$

下面，针对装备数据集中的装备名称实例分别采用 Jaro-Winkle、最小编辑距离、余弦距离等相似度算法和本文算法进行计算，得到的准确率、精确率、召回率和 F1($\beta=1$)结果如图 3 所示。

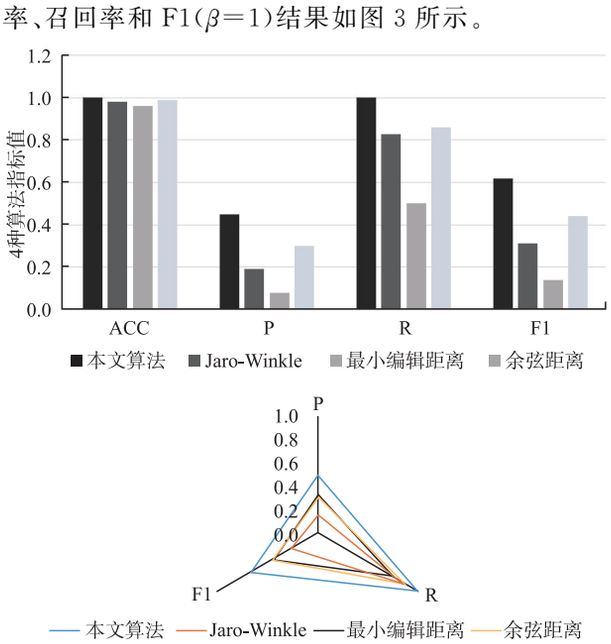

图 3 4 种算法评价指标综合对比

1)4 种算法的 ACC 值较为接近，主要原因是数据集中的反例远多于正例，所以 ACC 值参考意义不大；

2)各算法的 P 值都相对较低,表明符合匹配条件的反例较多,分析主要原因是部分装备名称相似度极高,横向比较,本文算法的 TP 和 FP 相对较少;

3)通过对结果的综合比较分析,本文算法的精确率、召回率、F1 值相较其它 3 种算法都有较好的表现。

对匹配成功和失败的数据样本进一步观察,可以总结出 3 点规律:一是装备名称包含的前后缀标识越多,本文算法的准确率越高,相较其它算法优势越明显;二是算法的有效性与数据质量有较大关联性,装备名称过于简单或存在歧义的数据样本无论在那种算法中都难以匹配成功;三是文中的所有基于字符的相似度算法都是纯文本分析,难以剔除极度相似的装备名称反例。

5　结语

本文在分析当前的主流的文本聚合算法的基础上,针对装备数据的特点,构建了装备数据的聚合模型并提供了相应的相似度算法,经过实验分析,该算法对装备数据适配性较高,能够一定程度上解决装备名称不一致带来的数据聚合难题,但如果想从源头上消除装备名称不一致现象仍需要在系统研发、数据采报等各个环节的工作中严格落实制度标准,确保装备字典的统一。本文是对装备数据聚合工作的初步探索,在未来的工作中可在以下 2 个方面做进一步研究:针对装备名称命名规范进行分析,寻求合适的分词方法以进一步提升聚合效率;构建专用的外部词典,尝试将字符距离相似度计算与语义相似计算相结合,以剔除极度相似的反例,进一步提升匹配的精确率。

参考文献

[1] 任连生.基于信息系统的体系作战能力概论[M].北京:军事科学出版社,2010:84-96.

[2] 李亢,李新明,刘东.多源异构装备数据集成研究综述[J].中国电子科学研究院学报,2015,10(2):162-168.

[3] 刘震,陈晶,郑建宾,等.中文短文本聚合模型研究[J].软件学报,2017,28(10):2674-2692.

[4] 刘兵,钱红琳.装备数据应用基本问题探析[J].装备学院学报,2015,26(1):107-110.

[5] 庄严,李国良,冯建华.知识库实体对齐技术综述[J].计算机研究与发展,2016,53(1):165-192.

[6] KHALID M A, JIJKOUN V, DE RIJKE M. The Impact of named Entity Normalization on Information Retrieval for Question Answering[C]//In Proc of the European Conf on Information Retrieval. Berlin, Heidelberg: Springer-Verlag, 2008: 705-710.

[7] BRIZAN D G, TANSEL A U. A Survey of Entity Resolution and Record Linkage Methodologies[J]. Communications of the IIMA,2006,6(3):41-50.

[8] LEE S, HWANG S. ARIA: Asymmetry Resistant Instance Alignment[C]//In Proc of the 28th AAAI Conference on Artificial Intelligence. Palo Alto: Association for the Advancement of Artificial Intelligence, 2014:94-100.

[9] YOU B, YAN Y S, SUN Y G, et al. Method of Information Content Evaluating Semantic Similarity on HowNet[J]. Computer Systems and Applications, 2013,22(1):129-133.

[10] PEDERSEN T, PATWARDHAN S, MICHELIZZI J. WordNet: Similarity: Measuring the Relatedness of concepts[C]//In Demonstration Papers at HLT-NAACL 2004. Palo Alto: Association for Computational Linguistics, 2004:38-41.

[11] KUMAR S, RANA J L, JAIN R C. Text Document Clustering Based on Phrase Similarity Using Affinity Propagation[J]. Int'l Journal of Computer Applications, 2013,68(10):38-44.

[12] JIANG H, HAN A Q, WANG M J, et al. Solution Algorithm of String Similarity Based on Improved Levenshtein Distance[J]. Computer Engineering, 2014,40(1):222-227.

[13] HERZOG T H, SCHEUREN F, WINKLER W E. Record linkage[J]. Wiley Interdisciplinary Reviews: Computational Statistics, 2010,2(5):535-543.

[14] GOMAA W H, FAHMY A A. A Survey of Text Similarity Approaches[J]. Int'l Journal of Computer Applications, 2013,68(13):13-18.

[15] ROUSSEAU R . The F-measure for Research Priority[J]. Journal of Data and Information Science, 2018, 3(1):17-21.

(编辑:韩茜)