石油科学通报
Petroleum Science Bulletin

语义相似度计算在内检测数据参数匹配中的应用

张河苇[1]，金剑[2]，董绍华[1*]，张来斌[1]，李宁[2]

1 中国石油大学(北京)机械与储运工程学院，北京 102249
2 中石油管道有限责任公司西部分公司，乌鲁木齐 830000
* 通信作者, shdong@cup.edu.cn

收稿日期: 2018−04−20
国家重点基础研究发展计划(2017YFC0805800) 和中石油管道有限责任公司西部分公司科研项目“管道完整性大数据架构模型及辅助决策分析模型研究与应用”(XG−JCGL−CX−KJXX−01−JL−03/201604) 联合资助

摘要　内检测数据对齐有助于提高内检测数据的利用率，目前国内外学者已初步建立内检测对齐流程。然而针对管道大数据背景下需匹配字段繁杂、中文字段描述多样等问题仍缺乏解决方案。本文采用中文语义相似度计算方法，计算各类字段与模板字段的相似度，确定其匹配度，可以从大量字段中选取匹配字段，实现不同来源内检测数据的对齐。本文在原有的基于同义词词林计算方法的基础上进行改进，并使用内检测报告中的实际字段进行计算，通过比对发现，本文改进的方法能够区分内检测报告中的不同字段，对多来源内检测数据对齐有较好的适用性。

关键词　语义相似度；内检测；数据对齐；同义词词林；长输管道

0　引言

管道内检测数据是维护管道运行的重要参考数据，通过内检测数据可以获得很多有价值的信息，是管道公司查找管道缺陷、进行管道修复的重要依据，因此针对同一管段往往会进行多轮内检测。然而，由于管道所处的环境以及检测过程中部分因素的影响，例如起始点不同、内检测器运行速度不同等，使得多次内检测数据无法完全对齐，降低了数据的利用水平，出现缺陷无法匹配等问题，甚至如果两次检测的检测商不同，则会进一步加剧这个情况。针对多轮内检测的比对问题，目前检测公司都是通过人工比对两次内检测数据，业务量巨大，而且对于管道运营商来说无法确定结果的真实性。

鉴于内检测数据比对的重要性，近期在内检测数据的比对理论方面，部分国内学者也进行了一些研究。王良军等[1]通过调研了解到国外的管道运营公司，例如DOW、BP、EnbrigeSingapore、Gas Company等，已有百余条管道开展了内检测数据比对工作。王良军等综述了内检测比对方法的研究现状，归纳出此项研究工作中的两个关键步骤为内检测里程数据对齐和内检测特征数据比对[1]。王丹丹等[2]提出在确定关键点对齐的前提下，以相对里程、时钟方位以及表面位置为核心参数的比对方法，并运用改进方法对海底管道的剩余强度和剩余寿命进行了评估。孙浩等[3]对内检测比对的流程进行详细叙述，包括关键点对齐和缺陷的活性判断方法，并以天然气管道为例进行方法验证，得到较好的计算效果，其限制条件主要为内检测数据须由同一检测承包商提供。杨贺[4]对比对中的关键流程(焊缝对齐、缺陷点识别)算法进行了设计，其限制条件为导入文件的格式必须与模板一致。

引用格式：张河苇, 金剑, 董绍华, 张来斌, 李宁. 语义相似度计算在内检测数据参数匹配中的应用. 石油科学通报, 2018, 04: 446−451
ZHANG Hewei, JIN Jian, DONG Shaohua, ZHANG Laibin, Li Ning. Application of semantic similarity calculation in parameter matching of detection data. Petroleum Science Bulletin, 2018, 04: 446-451. doi: 10.3969/j.issn.2096-1693.2018.04.040

现阶段内检测数据比对方法的基本流程已经确定，存在的问题主要是缺少快速匹配不同检测商提供的内检测报告的方法，该问题的存在限制了大数据背景下的数据对齐研究。通过语义相似度计算方法研究，有利于建立数据匹配字段的关联关系，实现数据的快速入库，为大数据技术的应用奠定基础。

1 基础理论

语义相似度计算是处理自然语言的重要研究内容，在信息检索、翻译等涉及到同义匹配等领域均有应用。目前绝大多数描述概念词语相似度的计算模型的基本思想是Dekang Lin从信息论的角度给出的如式1所示的理论[5]。含义为任意两个对象之间的相似度取决于它们之间的共性commonality和个性differences，共性越多，相似度越大；个性越多，相似度越小[6]。

$$sim(A,B)=\frac{\log\left(commen(A,B)\right)}{\log\left(description(A,B)\right)} \tag{1}$$

式(1)中的分母表示完整描述A，B所需的信息量大小，分子表示描述A，B共性部分所需的信息量大小，$sim(A，B)$表示A，B之间的语义相似度。

语义相似度计算的研究领域主要分为两大类[7]：一是依据某种世界知识来计算，主要是通过词典中概念结构关系(上下位关系、同位关系、整体－部分关系等)来计算相似度；二是利用大规模的语料库，利用统计学方法将上下文信息的概率分布作为词语语义相似度的度量。

本文研究的方法属于第一类。目前国外的语义研究词典主要包括WordNet[8]、FrameNet[9]、MindNet[10]等。国内的汉语语义研究词典主要为知网[11]、同义词词林[12]等。由于《同义词词林》的编排结构与国际研究常用的WordNet词典结构最为相似，该词典已逐渐成为汉语语义研究的重点，本文讨论的方法也是基于同义词词林建立的。

1.1 同义词词林

《同义词词林》是1983年由梅家驹等[12]编纂而成的。后来哈工大信息检索研究实验室根据人民日报语料库中词语出现的频率对其进行扩展并对词林的结构和编码进行了改进，形成一部具有汉语大词表的《哈工大信息检索研究室同义词词林扩展版》(《词林扩展版》)，共包含77 343词语。原版中只针对大类、中类、小类进行了编码，而《词林扩展版》形成了5层结构，同时将编码等级由三级扩充到了五级，划分为12个大类，95个中类，1428个小类，小类下方进一步划分为4026个词群和17 797个原子词群[13]。《同义词词林》扩展前后词典文件特征对比如表1所示。

同义词词林词典的5层结构如图1所示。上面四层的结点都代表抽象的类别，第5层的叶子结点表示具体的词条或义项[14]。对应5层结构设置了5层编码，第1层用大写英文字母表示；第2层用小写英文字母表示；第3层用二位十进制整数表示；第4层用大写英文字母表示；第5级用二位十进制整数表示。编码总长度为8位，结构具体如表2所示。

需要注意的是，第8位的标记有“=”、“#”、“@”3种。其中，“=”代表“相等”、“同义”；“#”代表“不等”、“同类”，表示属于同类，但是语义不同；“@ ”代表“自我封闭”、“独立”，它在词典中既没有同义词，也没有相关词[12]。

1.2 语义相似度计算方法

部分学者在基于同义词词林的语义相似度计算方法研究方面已取得一定的成果，认可度较高的有田久乐[15]和王汀[16]提出的算法。

表1 《同义词词林》扩展前后词典文件特征对比

Table 1 Comparison of dictionary file features before and after Synonym Word Forest expansion

特征	扩展前	扩展后
词语总数	53859	77343
大类数目	12	12
中类数目	94	97
小类数目	1428	1400
层数	3	5
编码长度	4	8

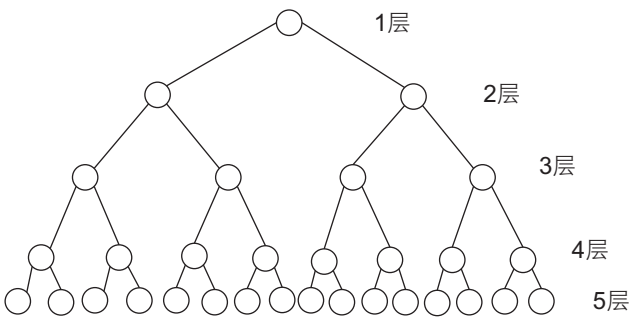

图1 同义词词林词典结构

Fig. 1 Synonym Word Forest dictionary structure

表 2 编码结构

Table 2 Coding structure

编码位	1	2	3 4	5	6 7	8
符号	大写英文字母	小写英文字母	两位十进制数	大写英文字母	两位十进制数	“=”、“#”、“@”
类别	大类	中类	小类	词群	原子词群	
层级	1 层	2 层	3 层	4 层	5 层	

1.2.1 田久乐算法

田久乐提出基于义项的语义距离来衡量词语的相似度[15]。假设两个义项 A，B 的相似度用 sim 表示。

(1)若两个义项不在同一棵树上

$$sim(A,B)=f \tag{2}$$

(2)若两个义项在同一颗树上

若在第 2 层分支，系数取 a，

$$sim(A,B)=1\times a\times\cos\left(n\times\frac{\pi}{180}\right)\left(\frac{n-k+1}{n}\right) \tag{3}$$

若在第 3 层分支，系数取 b，

$$sim(A,B)=1\times 1\times b\times\cos\left(n\times\frac{\pi}{180}\right)\left(\frac{n-k+1}{n}\right) \tag{4}$$

若在第 4 层分支，系数取 c，

$$sim(A,B)=1\times 1\times 1\times c\times\cos\left(n\times\frac{\pi}{180}\right)\left(\frac{n-k+1}{n}\right) \tag{5}$$

若在第 5 层分支，系数取 d，

$$sim(A,B)=1\times 1\times 1\times 1\times d\times\cos\left(n\times\frac{\pi}{180}\right)\left(\frac{n-k+1}{n}\right) \tag{6}$$

(3)若两个义项的编号相同，当末尾号为“=”时，相似度为 1；当末尾号为“#”时，直接把定义的系数赋给结果；当末尾号为“@”时，因代表在一个编号中只有一个词，所以不予考虑。

$\cos\left(n\times\frac{\pi}{180}\right)$ 表示调节参数，其作用是把相似度控制在[0,1]之间，n 代表分支层的节点总数；$(n-k+1)/n$ 表示控制参数，其作用是对结果精确化，n 代表分支层的节点总数，k 代表两个分支间的距离；在经过多次试验后，各系数初值设置为 a= 0.65，b=0.8，c= 0.9，d= 0.96，e= 0.5，f=0.1。

需要注意的是，针对有多个编码的词语，在计算词语相似度时，取最大值。

1.2.2 王汀算法

相较于田久乐提出的算法，王汀算法引入了概念相似度权重系数 $\lambda\times(L_i/|L|)$，L_i={1,2,3,4,5}，$|L|$ 表示集合 L 中的元素个数，恒等于 5。算法公式[16]如式 7 所示。

$$SIM_T(C_s,C_t)=\lambda\times\frac{L_i}{|L|}\times\cos\left(N_t\times\frac{\Pi}{180}\right)\times\left(\frac{N_t-D+1}{N_t}\right) \tag{7}$$

$\lambda\in(0,1)$，其取值不宜过高；N_t 为词元在第 i 层分支上的节点总数；D 为词元的编码距离；特别地，当概念的 5 层编码均相等且词林编码末位为“=”时，SIM_T 的取值为 1.0。

权重系数的引入使得不同层级的语义相似度区分更为明确。

2 基于内检测参数的语义相似度计算方法改进

使用前文介绍的两种方法进行实验验证发现，大部分的字段可以被区分开，然而部分字段的相似度计算差值较小，甚至无法区分，主要原因是未考虑路径对语义相似度的影响。由上文两种算法的公式可以看出田久乐算法仅设置了层级系数，王汀算法也只针对层级系数进行调节。本文通过增加路径权重对上述两种方法进行改进，改进后的公式如式 8 所示：

$$sim=\frac{L_i}{L}\cos\left(N\frac{\pi}{180}\right)weight\frac{N-K+1}{N} \tag{8}$$

式(8)中引入新的概念—路径权重 $weight$ 替代原参数 λ，目的是增大路径所在层级对相似度值计算结果的影响。取值参照表 3[17]中的设定值；L_i/L 为深度调节参数，L_i={1,2,3,4,5}，L=5；N 代表分支层的节点总数；K 代表两个义项的父节点的间距。路径、深度以及 N、K 的含义如图 2 所示，图中 K=3，N=6。特别地，若两个义项的编号相同，当末尾号为“=”时，认为相似度最大；当末尾号为“#”时，认为相似度最小；当末尾号为“@”时，因代表在一个编号中只有一个词，所以不予考虑。

表 3 路径权重设定值

Table 3 Setting value of path weight

层级	权重
1 层	0.5
2 层	1.5
3 层	4
4 层	6
5 层	8

3 案例

为了验证本文改进方法的有效性，选取管道缺陷描述字段中较难区分的模板字段：焊缝和沟槽，进行算法分析对比。将沟、坑痕、陷坑、槽子等几个描述词语与模板字段(焊缝和沟槽)通过两两计算语义相似度进行匹配。查询同义词词林[13]得到各字段的编码如表 4 所示。

采用田久乐算法得到的结果如表 5 所示，采用王汀算法得到的结果如表 6 所示。由计算结果可知，王汀方法在沟槽与槽子、焊缝与槽子的语义相似度计算中，差值为负数，未能成功匹配。

采用本文算法得到结果如表 7 所示，比较 3 种方法的差值增加量如表 8。对比可知，本文方法相对于其他两种方法非匹配字段的差值均有所增大，字段区分更为明显，并且能够区分其他方法难以区分的字段。

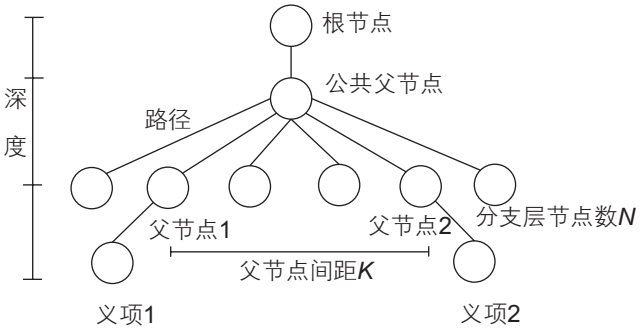

图 2 示意图

Fig. 2 Schematic

表 4 字段编码表

Table 4 Field coding table

	字段名称	字段编号	
模板字段	焊缝	Cb22B05@	
	沟槽	Bn15A01	
描述字段	沟	Be05C02	Bn08C01
	坑痕	Bg09B15#	
	陷坑	Bn13D01	
	槽子	Ba05A15#	

表 5 田久乐方法计算结果

Table 5 Calculation results of Tian's method

	沟		坑痕	陷坑	槽子
	Be05C02	Bn08C01	Bg09B15#	Bn13D01	Ba05A15#
沟槽Bn15A01	0.34	0.54	0.41	0.7	0.2
焊缝Cb22B05@	0.1	0.1	0.1	0.1	0.1
差值	0.24	0.44	0.31	0.6	0.1

表 6 王汀方法计算结果

Table 6 Calculation results of Wang's method

	沟		坑痕	陷坑	槽子
	Be05C02	Bn08C01	Bg09B15#	Bn13D01	Ba05A15#
沟槽Bn15A01	0.21	0.41	0.25	0.53	0.13
焊缝Cb22B05@	0.19	0.19	0.19	0.19	0.19
差值	0.02	0.22	0.06	0.34	−0.06

表 7 本文方法计算结果

Table 7 Calculation results of the improved method

	沟		坑痕	陷坑	槽子
	Be05C02	Bn08C01	Bg09B15#	Bn13D01	Ba05A15#
沟槽Bn15A01	0.316	1.644	0.38	2.1	0.188
焊缝Cb22B05@	0.064	0.064	0.064	0.064	0.064
差值	0.252	1.58	0.316	2.036	0.124

表 8 计算结果对比

Table 8 Calculation results comparison

田久乐法	本文方法	增加量	王汀法	本文方法	增加量
0.24	0.252	0.012	0.02	0.252	0.232
0.44	1.58	1.14	0.22	1.58	1.36
0.31	0.316	0.006	0.06	0.316	0.256
0.6	2.036	1.436	0.34	2.036	1.696
0.1	0.124	0.024	−0.06	0.124	0.186

4 结束语

管道行业数据容量已经累计到大数据级别，建立大数据库能够有效提高数据利用率，实现数据描述字段的自动匹配，能够为智能化数据导入提供便利，节省人力物力。本文结合语义相似度计算算法，从内检测字段入手，通过增加路径权重改进现有计算方法，使其适用于管道行业。与其他方法对比证明了本文改进方法的有效性。现阶段管道行业亟待建立管道大数据，字段匹配结合已有的数据对齐流程可实现多轮次数据的对齐，提高数据利用率的同时，能够为发掘管道缺陷和风险预测奠定基础。

参考文献

[1] 王良军, 李强, 梁菁嬿. 长输管道内检测数据比对国内外现状及发展趋势[J]. 油气储运, 2015, 34(03): 233− 236. [WANG L J, LI Q, LIANG Y. The current situation and development trend of testing data ratio in long distance pipeline[J]. Oil And Gas Storage and Transportation, 2015, 34 (03): 233−236.]

[2] 王丹丹, 林晓, 骆秀媛, 等. 海底管道两轮漏磁内检测数据的比对方法[J]. 船海工程, 2016, 45(03): 122−126, 130. [WANG D D, LIN X, LUO X AI, et al. Comparison method for detecting data of two-wheel leakage in submarine pipelines[J]. Marine Engineering, 2016, 45 (03): 122−126, 130.]

[3] 孙浩, 帅健. 长输管道内检测数据比对方法[J]. 油气储运, 2017, 36(07): 775−780, 794. [SUN H, SHUAI J. Method for detecting data in long distance pipeline[J]. Oil and Gas Storage and Transportation, 2017, 36 (07): 775−780, 794.]

[4] 杨贺. 油气管道内检测数据比对分析方法及应用[D]. 大庆: 东北石油大学, 2017. [YANG H. Method and application of detection data comparison in oil and gas pipelines [D]. Daqing: University of Northeastern Petroleum, 2017.]

[5] 张亮, 尹存燕, 陈家骏. 基于语义树的中文词语相似度计算与分析[J]. 中文信息学报, 2010, 24(6): 23− 29. [ZHANG L, YIN C Y, CHEN J J. Chinese words similarity calculation and analysis based on semantic tree[J]. Chinese Journal of Information, 2010, 24 (6): 23−29.]

[6] LIN D. An information-theoretic definition of similarity[C]//Icml. 1998, 98(1998): 296−304.

[7] 葛斌, 李芳芳, 郭丝路, 等. 基于知网的词汇语义相似度计算方法研究[J]. 计算机应用研究, 2010, 09: 3329−3333. [GE B, LI F F, GUO S L, et al. Research on lexical semantic similarity calculation method based on knowledge network[J]. Computer Application Research, 2010, 09: 3329−3333.]

[8] MILLER G A, FELLBAUM C. Semantic networks of English[J]. Cognition, 1991, 41(1−3): 197−229.

[9] BAKER C F, FILLMORE C J, LOWE J B. The berkeley framenet project[C]//Proceedings of the 17th international conference on Computational linguistics-Volume 1. Association for Computational Linguistics, 1998: 86−90.

[10] RICHARDSON S D, DOLAN W B, VANDERWENDE L. MindNet: Acquiring and structuring semantic information from text[C]// Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics-Volume 2. Association for Computational Linguistics, 1998: 1098−1102.

[11] 董振东. 语义关系的表达和知识系统的建造[J]. 语言文字应用, 1998, 03: 76− 83. [DONG Z D. The expression of semantic relation and the construction of knowledge system[J]. Language Application, 1998, 03: 76−83.]

[12] 梅家驹. 同义词词林[M]. 上海: 商務印書館, 1984. [MEI J J. Synonym Forest [M]. Shanghai: Business Printing Library, 1984]

[13] 刘丹丹, 彭成, 钱龙华, 周国栋.《同义词词林》在中文实体关系抽取中的作用[J]. 中文信息学报, 2014, 28(2): 91−99. [LIU D D, PENG C, QIAN L H, ZHOU G D. The role of Thesaurus in Chinese entity relation extraction[J]. Chinese Journal of Information, 2014, 28 (2): 91−99.]

[14] 梅立军, 周强, 臧路, 等. 知网与同义词词林的信息融合研究[J]. 中文信息学报, 2005, 19(1): 63− 70. [MEI L J, ZHOU Q, ZANG L,

et al. The research of information fusion of knowledge net and Synonym Word Forest[J]. Chinese Journal of Information, 2005, 19 (1): 63−70.]

[15] 田久乐, 赵蔚. 基于同义词词林的词语相似度计算方法 [J]. 吉林大学学报: 信息科学版, 2010 (6): 602−608. [TIAN J L, ZHAO W. Word similarity calculation method based on thesaurus[J]. Journal of Jilin University: Information Science edition, 2010 (6): 602−608.]

[16] 王汀, 高迎, 刘经纬. 一种面向中文本体模式的本体对齐框架 [J]. 数据分析与知识发现, 2017, 1(2): 47−57. [WANG T, GAO Y, LIU J W. An ontology alignment framework for Chinese ontology mode[J]. Data analysis and knowledge discovery, 2017, 1 (2): 47−57.]

[17] 陈宏朝, 李飞, 朱新华, 等. 基于路径与深度的同义词词林词语相似度计算 [J]. 中文信息学报, 2016, 30(5): 80−88. [CHEN H C, LI F, ZHU X H, et al. Similarity calculation of thesaurus words based on path and depth[J]. Chinese Journal of Information, 2016, 30 (5): 80−88.]

Application of semantic similarity calculation in parameter matching of detection data

ZHANG Hewei[1], JIN Jian[2], DONG Shaohua[1], ZHANG Laibin[1], LI Ning[2]

1 School of Mechanical and Transportation Engineering, China University of Petroleum -Beijing, Beijing 102249, China
2 China Petroleum Pipeline Co., Ltd. West Branch, Urumqi 830000, China

Abstract The alignment of inline inspection datasets can help to improve the utilization rate of the data. At present, domestic and foreign scholars have preliminarily established the alignment method. However, there is still a lack of solutions to the complexity and the diversity of Chinese characters, which are used in the inline inspection reports. Here the method of Chinese semantic similarity calculation was used to determine the matching degree between fields, select the matched fields from a large number of fields and achieve the data alignment between different testing companies. This method is improved based on Synonym Forest, and the actual fields from the inline inspection test reports are used. The improved method can distinguish the different fields and has good applicability to the multiple inspection data alignment.

Keywords semantic similarity; inline inspection; data alignment; Synonym Forest; long distance pipeline

doi: 10.3969/j.issn.2096-1693.2018.04.040

(编辑 马桂霞)