doi:10.3969/j.issn.1672-4348.2017.04.009

基于改进 KNN 的案例匹配模块的设计与实现

谢开池，薛醒思

（福建工程学院 信息科学与工程学院，福建 福州 350118）

摘要：为了提高 KNN 检索策略的检索效率和检索结果的质量，提出一种改进的 KNN 检索策略。在引入图书馆领域本体和概念语义相似度度量技术的前提下，利用句法结构筛选不合理的案例以降低计算规模，从而提高案例的检索质量和效率，利用改进的微粒群算法优化概念语义相似度度量技术中的组合参数以提高 KNN 检索的结果质量。实验数据采用福州晓锋科技信息咨询有限公司提供的图书馆参考咨询测试数据。实验结果表明，相比于传统 KNN 和基于传统 PSO 的改进 KNN 方案有效地提高了案例匹配结果的查全率和查准率。

关键词：案例推理；KNN；微粒群算法

中图分类号：TP182　　**文献标志码**：A　　**文章编号**：1672-4348(2017)04-0349-09

Design and implementation of a case matching module based on improved KNN

Xie Kaichi, Xue Xingsi

(College of Information Science and Engineering, Fujian University of Technology, Fuzhou 350118, China)

Abstract: To improve the efficiency and quality of case retrieval, an improved KNN retrieval strategy was proposed. By introducing library domain ontology and concept semantic similarity measurement technology, cases' syntactic structure was employed to filter out the unreasonable cases to reduce the computation amount (search space) and improve the case retrieval (alignment's) quality. Then, an improved particle swarm algorithm was presented to determine the optimal aggregating parameters in the similarity measure technologies to improve the case alignment's quality. In the experiment, the testing cases were from Fuzhou Xiaofeng Science and Technology Information Consulting Ltd., Co,. The experimental results show that compared with the traditional KNN and the traditional PSO-based KNN, the proposal can significantly improve the case alignment's quality in terms of both recall and precision.

Keywords: case based reasoning; K nearest neighbourhood; particle swarm algorithm

图书馆虚拟咨询参考系统是一类智能决策支持系统(intelligent decision support systems,简称 IDSS)[1],该系统利用基于知识库和案例库的推理技术确定用户决策所需的信息。案例推理技术[2](case based reasoning,简称 CBR)是 IDSS 的核心技术之一,其目的是依据给定的决策问题和决策环境的特征描述,快速而有效地确定案例库中对求解的问题最有帮助的案例[3]。在案例推理技术中,如何制定一种高效的案例检索策略是目前亟需解决的关键问题[4]。针对该问题,李小展[5]提出了分阶段的最近邻策略,利用特征加权的产 K nearest neighbourhood(KNN)逐步缩小案例检索范围,以此提高医疗辅助诊断系统的检索效率;费玉莲[6]采用计算概念之间的相关系数的

收稿日期：2017-05-26
基金项目：国家级大学生创新创业训练计划项目(201610388024)
通讯作者：薛醒思(1981-),男,福建福清人,副教授,博士,研究方向:智能计算和本体技术的研究与应用。

语义距离求得概念间的相似度,再利用 KNN 求得源案例和目标案例相似度。但这些检索策略既没有全面考虑影响概念语义相似度的相关因素(语义关系、语义距离、节点深度和节点密度),且相似度匹配中参数仍存在人为因素,从而影响检索结果的质量。针对上述过程,提出一种基于图书馆领域本体知识库的 KNN 案例检索策略。首先,在引入图书馆领域本体[7]的基础上,利用本体的概念层次树结构对案例问句进行"主-谓-宾"三元组语义标注的案例标注,利用案例问句的"主-谓-宾"句式结构过滤案例库中不符合的案例集,减少候选案例集的规模,提高基于 KNN 的 CBR 技术的效率和结果质量。其次,通过改进的微粒群优化算法确定三类相似度度量技术最优的集成权重、K 值及相似度阈值,并在匹配过程中综合考虑 3 类基于本体概念层次树的相似度度量技术,提高了案例检索的查准率和查全率。

1 相关概念与技术

定义 1 本体(Ontology):本体是一个共享概念化模型的形式化和显式的说明规范[8],其核心是本体能够解释计算机语义信息。文中本体可表示为一个三元组 $O=(C,R,A)$。其中:C 表示概念的集合,即某一领域的概念范畴;R 表示概念间的关系,即该领域概念之间的关联;A 表示公理集,即关于被建模领域中真假的论述。

定义 2 本体概念层次树:给定本体 O,本体概念层次树是一个三元组 $T=(N,R,F)$。其中:$N=\{n_1,n_2,\cdots,n_e\}$ 是本体概念层次树中个概念节点的集合,$n_i(i=1,2,\cdots,e)$ 对应 O 中的概念 $c_i(i=1,2,\cdots,|C|)$;$R=(n_i,n_j)\ i,j\in\{1,2,\cdots,e\}$ 是本体概念层次树中概念节点 n_i 和 n_j 之间的边集合,表示对应概念 c_i 和 c_j 之间的 is-a 关系;$F=\{F_1,F_2,\cdots\}$ 是本体概念层次树的层的集合。其中:令 $\mathrm{path}(n_i)$ 表示概念节点 n_i 到根节点的最短路径长度,$F_j=\{n_i\mid n_i\subseteq N,\mathrm{path}(n_i)=j,i\in[1,e],j=1,2,\cdots\}$。

定义 3 图书馆参考咨询案例:图书馆参考咨询案例是一个五元组 case = (caseId,question,predicate,keyword,answer),其中 caseId 表示图书馆参考咨询案例的编号;question 表示该案例的问题;predicate 表示该案例的谓语;keyword = (subject,object) 表示该案例的关键词(分别是主语和谓语);answer 表示该案例的问题答案。

定义 4 图书馆参考咨询案例库:图书馆参考咨询案例库是指一种将已存在的图书馆参考咨询案例以一定的索引方式存储的知识库,表示为 $\mathrm{Case}=\{\mathrm{case}_1,\mathrm{case}_2,\cdots,\mathrm{case}_z\}$,$\mathrm{case}_z$ 表示案例库中的第 Z 个历史案例(history case,简称 HC),Z 表示图书馆参考咨询案例的个数。在本文的工作中,尚未保存在案例库中的新案例称为目标案例(new case,简称 NC),同时将图书馆参考咨询案例库分成训练集和测试集两部分,训练集表示为 $\mathrm{TrainCase}=\{\mathrm{case}_i\mid\mathrm{case}_i\subseteq\mathrm{Case},i\in[1,\mathrm{num}_{\mathrm{train}}],\mathrm{num}_{\mathrm{train}}<Z\}$,测试集 $\mathrm{TestCase}=\{\mathrm{case}_j\mid\mathrm{case}_j\subseteq_{\mathrm{Case}}\mathrm{TrainCase},j\in[1,Z-\mathrm{num}_{\mathrm{train}}]\}$,其中 $\mathrm{num}_{\mathrm{train}}$ 表示训练集中案例的个数。

定义 5 图书馆参考咨询案例检索词集合:给定 num_q 个检索词,图书馆参考咨询案例检索词集合定义如下:$Q=\{q_1,q_2,..,q_{num\_q}\}$,其中 $q_i=(\mathrm{subject}_i,\mathrm{predicate}_i,\mathrm{object}_i)$,$i=1,2,\cdots,\mathrm{num\_}q$ 表示第 i 个检索词,$\mathrm{subject}_i$、$\mathrm{predicate}_i$ 和 object_i 分别表示检索词的主语、谓语和宾语。

定义 6 图书馆参考咨询案例匹配结果质量度量技术:给定案例参考匹配结果 Refer 和案例匹配结果 ReturnCase(指案例匹配的结果),图书馆参考咨询案例匹配结果的查全率 $r(\mathrm{ReturnCase},\mathrm{Refer})$、查准率 $p(\mathrm{ReturnCase},\mathrm{Refer})$、度量 $f(\mathrm{ReturnCase},\mathrm{Refer})$ 分别定义如下[9]:

$$r(\mathrm{ReturnCase},\mathrm{Refer})=\frac{|\ \mathrm{Refer}\cap\mathrm{ReturnCase}\ |}{|\ \mathrm{Refer}\ |}\tag{1}$$

$$p(\mathrm{ReturnCase},\mathrm{Refer})=\frac{|\ \mathrm{Refer}\cap\mathrm{ReturnCase}\ |}{|\ \mathrm{ReturnCase}\ |}\tag{2}$$

$$f(\mathrm{ReturnCase},\mathrm{Refer})=\frac{2\times r(\mathrm{ReturnCase},\mathrm{Refer})\times p(\mathrm{ReturnCase},\mathrm{Refer})}{r(\mathrm{ReturnCase},\mathrm{Refer})+p(\mathrm{ReturnCase},\mathrm{Refer})}\tag{3}$$

公式(1)-(3)中:$|\ \mathrm{Refer}\cap\mathrm{ReturnCase}\ |$ 表示匹配结果参考匹配结果的交集个数;$|\ \mathrm{Refer}\ |$ 表示参考匹配结果的个数;$|\ \mathrm{ReturnCase}\ |$ 表示匹配结果的个数。

定义 7 概念语义相似度和案例相似度:概

念语义相似度是概念间可以互相替换的程度[10]。通过综合考虑两个概念的语义距离、节点深度和节点密度[11]来度量它们之间的概念语义相似度。

(1)语义距离是指在本体概念层次树中，概念 n_i 和 n_j 到两者最近共同父节点的路径长度 distance(n_i,n_j)。如果 distance(n_i,n_j) 越小，即概念的语义距离越小，其相似度越大；反之则相似度越小。

(2)节点深度是指在本体概念层次树中某一概念节点 n_i 所处层次集 F_{depth}(depth = 1,2,…) 到根节点的路径长度 depth。若两个节点 n_i 和 n_j 的层次数之和 depth(n_i) + depth(n_j) 越大，对应的概念之间的相似度则越大；若两个节点的层次数之差 | depth(n_i) − depth(n_j) | 越小，对应概念之间的相似度则越大。

(3)节点密度是指在本体概念层次树中，概念节点 n_i 和 n_j 的共同父节点拥有的子节点密度。则二者的节点密度定义如下：

$$\text{density}(n_i,n_j)=\frac{p}{q} \tag{4}$$

其中：p 是指以概念节点 n_i 和 n_j 的共同父节点为根节点的子树所包含的除根节点以外的节点个数；q 是指节点 n_i、n_j 和它们共同父节点之间所构成树的最大层次差。

在研究中，给出两个概念节点 n_i 和 n_j，它们之间的概念语义相似度度量公式定义如下[12]：

$$\begin{aligned}\text{simconcept}(n_i,n_j)=&\left[\frac{a}{\text{distance}(n_i,n_j)+a}\right]^{x_1}\times\\&\left[\frac{\text{depth}(n_i)+\text{depth}(n_j)}{|\text{depth}(n_i)-\text{depth}(n_j)|+1}\right]^{x_2}\times\\&\left[\frac{1}{\text{density}(n_i,n_j)}\right]^{x_3}\end{aligned} \tag{5}$$

其中：depth(n_i) 和 depth(n_j) 分别表示 n_i 和 n_j 的概念节点深度；density(n_i,n_j) 表示 n_i 和 n_j 的节点密度；x_1、x_2 和 x_3 分别表示基于语义距离、节点深度和节点密度的相似度度量技术的集成权重且 $\sum_{i=1}^{3}x_i=1, x_i\in(0,1)$。

给定历史案例 HC 和目标案例 NC，二者的案例相似度定义如下：

$$\text{simcase}(HC,NC)=\text{simallconcept}/4 \tag{6}$$

式中，

$$\begin{aligned}\text{simallconcept}=&\max(\text{simconcept}(hc_{\text{subject}},nc_{\text{subject}}),\text{simconcept}(hc_{\text{subject}},nc_{\text{object}}))+\\&\max(\text{simconcept}(hc_{\text{object}},nc_{\text{subject}}),\text{simconcept}(hc_{\text{object}},nc_{\text{object}}))+\\&\max(\text{simconcept}(nc_{\text{subject}},hc_{\text{subject}}),\text{simconcept}(nc_{\text{subject}},hc_{\text{object}}))+\\&\max(\text{simconcept}(nc_{\text{object}},hc_{\text{subject}}),\text{simconcept}(nc_{\text{object}},hc_{\text{object}}))\end{aligned} \tag{7}$$

其中：hc_{subject} 表示历史案例中的主语；hc_{object} 表示历史案例中的宾语；nc_{subject} 表示目标案例的主语；nc_{object} 表示目标案例的宾语；simconcept 表示主宾间相似度，用公式(5)计算。

2　基于本体和微粒群算法的改进 KNN 技术

基于改进 KNN 的案例检索策略分两个步骤：基于改进 PSO 的参数训练步骤和案例测试步骤。已知检索词集合的前提下，首先利用改进 PSO 算法和案例训练集 TrainCase 寻优得到参数集合 A；然后基于给定的案例测试集 TestCase，利用训练的参数集合 A 和 TestCase 中某一案例检索，得到该案例的匹配案例集合。该技术的框架图如图 1 所示。

2.1　案例检索策略中确定参数集合的单目标优化模型

在案例检索过程中，如何确定 KNN 检索策略中由相似度权重、K 值和阈值组成的最优参数集合，使得测试集中所有案例的检索结果质量最优是案例检索策略中的关键问题。基于明显的观察结果，结果案例的数量和结果案例间的相似度同案例匹配结果质量相关，因此针对该问题提出的单目标优化模型数学形式如下：

$$\begin{cases}\max f(X)\\ X=(x_1,x_2,\cdots,x_n)^{\mathrm{T}}\\ \text{s.t.}\,x_i\in[0,1],i=1,2,\cdots,n-2\\ \sum_{i=1}^{n-2}x_i=1\\ x_{n-1}=1,2,\cdots\\ x_n\in[0,1]\end{cases} \tag{8}$$

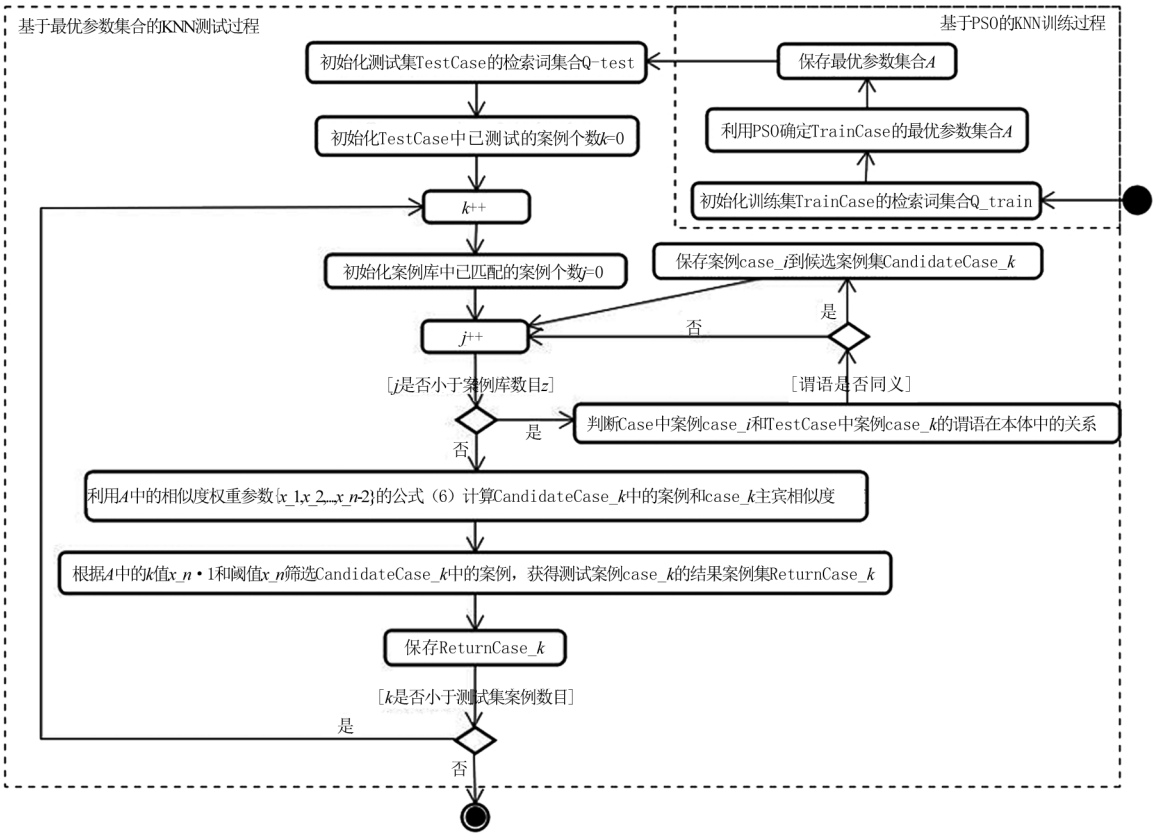

图 1 基于改进 KNN 的案例检索策略框架图

Fig.1 The framework of improved KNN-based case retrieval strategy

其中：$\boldsymbol{X}$ 是参数集合向量 $(x_1,x_2,\cdots,x_n)^{\mathrm{T}}$，$x_i(i=1,2,\cdots,n-2)$ 是概念相似度影响因素 i 的权重，x_{n-1} 是 KNN 策略中 K 值，x_n 是用于过滤同目标案例相似度值太低的源案例的阈值；$f(X)$ 是训练集中所有案例匹配结果的平均质量，其度量函数定义如下：

$$f(X)=\frac{\mathrm{value}_1+\mathrm{value}_2+\cdots+\mathrm{value}_{\mathrm{num}_{\mathrm{train}}}}{\mathrm{num}_{\mathrm{train}}} \tag{9}$$

式中：$\mathrm{value}_l(l=1,2,\cdots,\mathrm{num}_{\mathrm{train}})$ 表示训练集中案例 case_l 的评价值，其计算公式如下：

$$\mathrm{value}_l(X,NC)=0.5\times\frac{1}{\mathrm{num}_{\mathrm{return}}}+0.5\times\frac{\left|\bigcup_{i=1}^{\mathrm{num}_{\mathrm{return}}}\mathrm{case}_i\cap NC\right|}{|NC|} \tag{10}$$

其中：$\mathrm{num}_{\mathrm{return}}=|\mathrm{ReturnCase}|$ 表示结果案例集中案例的数量。设 $h:\{x_1,x_2,\cdots,x_{n-2}\}\rightarrow\mathrm{simcase}$ 是参数集合 $\{x_1,x_2,\cdots,x_{n-2}\}$ 到 simcase 的映射，设 $g:\{\mathrm{simcase},x_{n-1},x_n\}\rightarrow\mathrm{ReturnCase}$ 是参数集合 $\{\mathrm{simcase},x_{n-1},x_n\}$ 到 ReturnCase 的映射，则 ReturnCase 可定义为：$\mathrm{ReturnCase}=g(h(x_1,x_2,\cdots,x_{n-2}),x_{n-1},x_n)$；$\mathrm{case}_i$ 表示结果案例集 ReturnCase 中的案例；$\left|\bigcup_{i=1}^{\mathrm{num}_{\mathrm{return}}}\mathrm{case}_i\cap NC\right|$ 表示结果案例与目标案例的共同信息；$|NC|$ 表示目标案例的所有信息。

2.2 改进的自适应惯性权重的微粒群优化算法

由于求解的优化问题是一类复杂的非线性优化问题（目标函数不可微，且拥有大量的局部最优解），考虑到 PSO 在诸多求解复杂非线性优化问题中的成功应用[13-15]，使用 PSO 来确定检索策略中的参数集合。采用的是实数编码方案。群体中的每一个个体是一个长度为 n 的一维数组，前 n−2 个元素记为 $N_1N_2\cdots N_{n-2}$，其中 $\sum_{i=1}^{n-2}N_i=1$，$N_i\in[0,1]$，$i\in\{1,2,\cdots,n-2\}$，N_i 表示相似度度量技术中第 i 个影响因素的权重。个体编码的最后两位分别表示 KNN 中的 $K\in\{1,2,\cdots\}$ 和用于

过滤相似度计算结果的阈值 threshold ∈ [0,1]。

针对高维复杂函数优化，标准 PSO 算法存在收敛速度慢、易陷入局部最优的缺陷[16-18]。针对该缺陷，在求解最大化问题过程中的改进策略如下：设粒子群规模为 n，第 k 代粒子 x_i 的适应度值为 $f^{(k)}(x_i)$，第 $k-1$ 代粒子 x_i 的适应度值为 $f^{(k-1)}(x_i)$，第 k 代所有粒子的适应度平均值为 $f^{(k)}\text{xavg} = \dfrac{\sum_{i=1}^{n} f^{(k)}(x_i)}{n}$，第 k-1 代种群的全局最优适应度值为 $f^{(k-1)}\text{gbfitbest}$，第 k 代种群的全局最优适应度值为 $f^{(k)}\text{gbfitbest}$，若满足 $f^{(k)}\text{xavg} \leqslant f^{(k)}(x_i)$，则求取满足该条件粒子的平均适应度值为 $f^{(k)}\text{frontxavg}$；若满足 $f^{(k)}\text{xavg} > f^{(k)}(x_i)$，则求取满则该条件粒子的平均适应度值为 $f^{(k)}\text{behindxavg}$。

在设置了以上变量后，粒子的惯性权重依据以下原则赋值：

(1) 若 $f^{(k)}(x_i) > f^{(k)}\text{frontxavg}$，说明该粒子属于种群中优秀的粒子，该粒子的下一取值要趋近于收敛，w 取值范围内的最小值 0.4。

(2) 若 $f^{(k)}(x_i) < f^{(k)}\text{behindxavg}$，说明该粒子属于种群中较差的粒子，此时应该增加其全局寻优的能力，w 取值范围内的最大值 0.9。

(3) 若 $f^{(k)}\text{behindxavg} \leqslant f^{(k)}(x_i) \leqslant f^{(k)}\text{frontxavg}$，说明该粒子处于中等水平，正在逐步寻优，因此其取值按如下改进的公式计算：

$$w = w_{\min} + (w_{\max} - w_{\min}) \times \cos\left(\pi \times \frac{k}{\text{iter}_{\max}}\right) \times \frac{\min(f^{(k-1)}\text{gbfitbest}, f^{(k)}\text{gbfitbest})}{\max(f^{(k-1)}\text{gbfitbest}, f^{(k)}\text{gbfitbest})} \times \frac{\min(f^{(k)}\text{gbfitbest}, f^{(k)}\text{xavg})}{\max(f^{(k)}\text{gbfitbest}, f^{(k)}\text{xavg})} \tag{11}$$

其中：$w_{\max}$，$w_{\min}$ 分别为初始惯性权重和终止惯性权重，本文 $w_{\max}$ 取值为 0.9，$w_{\min}$ 取值为 0.4，k 为当前迭代次数，$\text{iter}_{\max}$ 为最大迭代次数。

通过以上的改进，防止当前迭代中粒子适应度变差的粒子引导继续向更差的方向移动，有效的降低了无效迭代的次数。同时，收敛速度加快，结果更稳定。改进的自适应惯性权重的 PSO 算法流程如图 2。

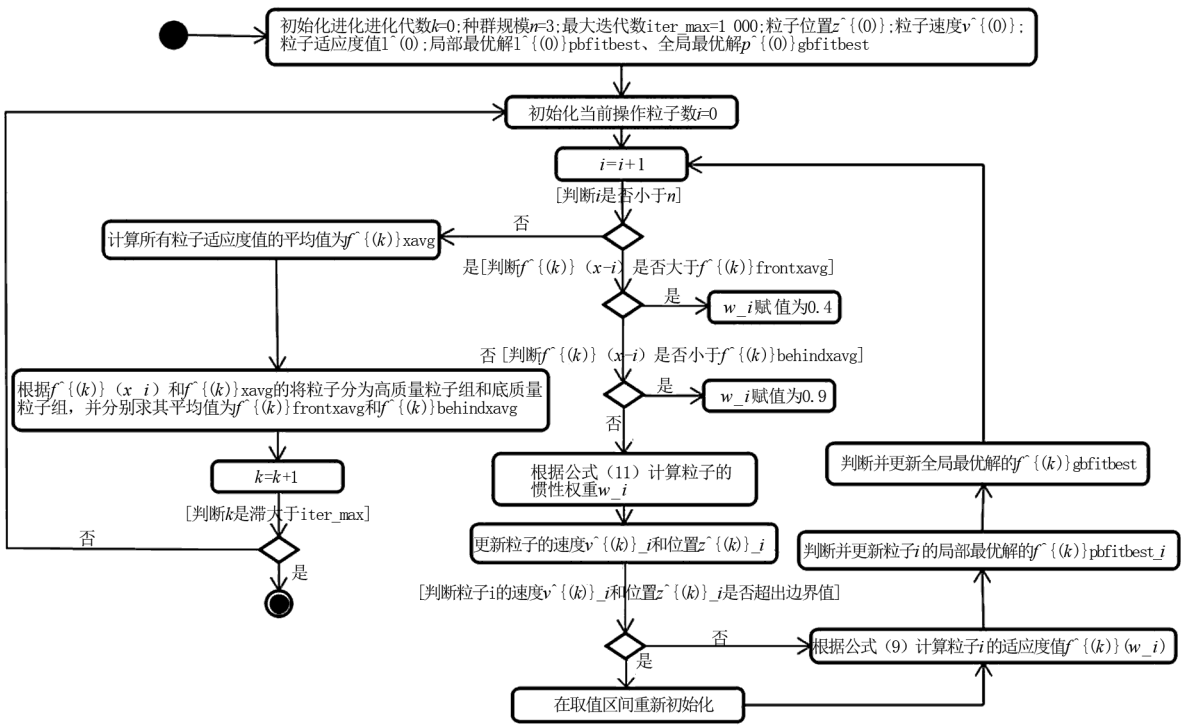

图 2　改进的 PSO 算法流程图

Fig.2　Comparison of the matching result's quality by f-measure among three schemes

3 实验结果与分析

选用的测试数据集(图书馆参考咨询数据)是由福州晓峰科技信息咨询有限公司提供的图书馆领域的相关知识。在案例的句法结构信息中,动词是其他成分(主语,宾语等名词)的基础,能直接影响案例匹配的正确率[19]。依据上述内容将测试数据分为两类来评价案例匹配模块的结果质量:(1)测试案例仅含有一个动词;(2)测试案例有两个及以上的动词。

每个测试用例由目标案例、检索词集合及参考匹配结果组成,分别表示用户咨询问题、经过分词扩展模块后的检索词集合和由专家确定的标准匹配结果。表 1 给出了本文测试用例的详细描述。

表 1 测试用例描述

Tab.1 The description of test cases

编号	目标案例	检索词集合	参考匹配结果
101	本校读者可以从其他图书馆借书吗?需要办理哪些手续?	(#,借,书)(读者,借,#)(读者,借,书)(学生,借,#)(学生,借,图书)(读者,办理,#)(#,办理,手续)(读者,办理,手续)	(本校读者可以从其它图书馆借书吗?)(本校读者从其它图书馆借书需要办理哪些手续?)(本校读者可以从其它图书馆借书吗,需要办理哪些手续?)(本校读者可以委托他人代办借书手续吗?)(请问我校学生如何到其它院校图书馆借书?)(如果 IC 卡丢了,可不可以到图书馆借书?)(国家图书馆所有的图书我们都可以借吗?)
102	本校教师借阅图书有特殊规定吗?	(#,借阅,图书)(教师,有,#)(#,有,规定)(图书,有,规定)(教师,有,规定)	(本校教师借阅图书有特殊规定吗?)(如果 IC 卡丢了,可不可以到图书馆借书?)(国家图书馆所有的图书都可以借吗?)
103	如果 IC 卡丢了,可不可以到图书馆阅览图书?	(#,阅览,图书)(#,丢,IC 卡)(#,丢,图书)	(如果 IC 卡丢了,可不可以到图书馆阅览图书?)(如果 IC 卡丢了,可不可以到图书馆借书?)(借书证丢失了怎么办?)
104	请问期刊 SAMPE J.在哪些图书馆有收藏?	(图书馆,收藏,#)(#,收藏,期刊)(图书馆,收藏,期刊)(图书情报机构,收藏,#)(图书情报机构,收藏,核心期刊)(#,有,期刊)(图书馆,有,期刊)(图书馆,有,#)	(《莆田永春宁德宋氏联谱》哪几家图书情报机构有收藏,有电子全文么?)(如何判断图书馆是否收藏 1966 年以前的某种西文期刊?)(请问期刊 SAMPE J.在哪些图书馆有收藏?)(请问图书馆现刊中,有多少种中文期刊?)(请问图书馆现刊中,有多少种外文期刊?)
105	清华学生通过馆际互借可以预约北大图书馆的图书吗?	(学生,预约,#)(#,馆际互借,图书)(#,预约,图书)(学生,预约,图书)(学生,馆际互借,图书)	(北图外文新书阅览室的书能馆际互借吗?)(如何知道预约的图书他人已还?)(哪些图书不可以预约?)(不到图书馆可以预约图书吗?)
201	请问我毕业后还可以使用图书馆吗?	(#,使用,图书馆)	(请问我毕业后还可以使用图书馆吗?)
202	国家图书馆所有图书我们可以借吗?	(#,借,图书)	(国家图书馆所有的图书我们都可以借吗?)
203	ISBN、ISSN 有什么含义?	(#,有,ISBN)(#,有,含义)(ISBN,有,含义)(#,有,国标号)(#,有,图书条码号)(图书条码号,有,类型)(图书条码号,有,含义)	(ISBN、ISSN 有什么含义?)(图书条码号、索书号有什么含义?)

续表

编号	目标案例	检索词集合	参考匹配结果
204	如何查询外文引文索引文献的“收录号”？	(#,查询,文献)(#,查询,收录号)	(如何查询外文引文索引文献的“收录号”?)(如何查找国内外标准文献?)(如何找到与课题相关的文献?)
205	图书条码号、索书号有什么含义？	(#,有,图书条码号)(#,有,含义)(图书条码号,有,含义)(#,有,索书号)(索书号,有,含义)	图书条码号、索书号有什么含义？

改进 PSO 的算法参数如下：

(1)基于本体的概念语义相似度度量的调节参数 a 用于调节概念相似度值，a 取值越大概念相似度值语义距离趋近于 1 的速度越快[20]，取值为 1 时相似度效果最好。

(2)粒子群种群规模及最大进化代数的设定与问题的规模成正比，设定值偏小会影响结果的质量，设定值偏大影响结果运行的效率。种群规模建议范围为[5,20]，最大进化代数建议范围为[500,2 000]。由于本文的问题规模不大(仅有 5 个参数需要确定)，因此种群规模及最大进化代数分别设置为 10 个个体和 1 000 次迭代。

(3) $w_{\max}$ 是惯性权重 w 的初始值，$w_{\min}$ 为粒子进化到最大迭代数的惯性权重值，当 $w_{\max}=0.9$，$w_{\min}=0.4$ 时优化问题取得最好的效果[21]，因此本文取值为 $w_{\max}=0.9$，$w_{\min}=0.4$。

(4)学习因子 c_1，c_2 分别表示粒子个体向局部最优和全局最优位置移动的能力，一般设为相同的值，常见的设定为 2[22]。

(5)随机因子 r_1，r_2 是也是影响粒子“自我学习”和“社会学习”的能力，一般取值为[0,1]之间的随机数[23]。

为了检验本方案的匹配速率和匹配结果质量，本文同标准的 KNN 检索策略方案进行比较。KNN 也采用本体提出的基于本体的概念语义距离相似度度量。KNN 中 K 初始值由案例库中历史案例的数量[24]确定：$K=\frac{\sqrt{HC}}{2}$(若 K 计算为非整数，则向上取整)。本文计算 K 初始值为 8。然后，通过交叉验证证明本文问题模型中 K 取值为方案 3 结果质量最好。

表 2 中的数据是基于句法结构过滤的 KNN

表 2　通过 f 度量值比较 3 种方案的匹配结果质量

Tab.2　Comparison of matching result's quality by f-measure among three schemes

编号	方案 1	方案 2	方案 3	f 度量值的改进程度/%(方案 1 与方案 3)	f 度量值的改进程度/%(方案 2 和方案 3)
101	0.33(0.20, 1.00)	0.36(1.00, 0.22)	0.77(1.00, 0.63)	133.33	113.89
102	0.33(0.20, 1.00)	0.41(1.00, 0.26)	0.77(1.00, 0.63)	133.33	87.80
103	0.80(0.67, 1.00)	0.13(1.00, 0.07)	1.00(1.00, 1.00)	25.00	669.23
104	0.40(0.25, 1.00)	0.47(1.00, 0.31)	0.94(1.00, 0.89)	135.00	100.00
105	0.57(0.40, 1.00)	0.26(1.00, 0.15)	0.91(1.00, 0.83)	59.65	250.00
201	1.00(1.00, 1.00)	0.13(1.00, 0.07)	1.00(1.00, 1.00)	0.00	669.23
202	0.33(0.20, 1.00)	0.26(1.00, 0.15)	0.72(1.00, 0.56)	115.15	176.92
203	1.00(1.00, 1.00)	0.80(1.00, 0.67)	1.00(1.00, 1.00)	0.00	25.00
204	0.80(0.67, 1.00)	0.55(1.00, 0.38)	1.00(1.00, 1.00)	25.00	81.82
205	0.67(0.50, 1.00)	0.80(1.00, 0.67)	1.00(1.00, 1.00)	49.25	25.00
均值	0.68(0.51, 1.00)	0.46(1.00, 0.30)	0.92(1.00, 0.85)	67.57	219.89

(方案 1)、基于改进 PSO 的 KNN(方案 2)及基于句法结构和改进 PSO 的 KNN(方案 3)3 种方法的匹配结果的 f 度量(查全率,查准率),其中本文的方法是在每个测试用例上独立运行 10 次后的平均 f 度量值。从表 2 可以看出,方法 3 的匹配结果的 f 度量值都明显优于方案 1 和方案 2,说明基于改进 PSO 的 KNN 检索策略明显优于标准的 KNN 策略。此外,从方案 3 的 f 值比对中可以看出所有测试案例都远远高于方案 2(至少高出 25%),该数据证明利用句法结构(动词)过滤候选案例集可以很大程度的提高匹配的质量。

图 3 通过比较方案 2 和方案 3 得出有无句法结构过滤策略对所有案例测试独立运行 10 次后的平均运行时间(时间单位为 ms)影响,从该图 3 可以看出经过句法结构过滤案例候选集的方案 3 的大部分远远低于未过滤案例候选集的方案 2。本文测试案例库的规模为 233 个案例,若是案例库的规模更大,未预先经过句法结构过滤的方案其运行速度将非常慢,从而影响算法的效率。图 4 通过比较标准 PSO 和本文改进的 PSO 进化代数与独立运行 10 次后的平均适应度值的变化关系。从该图 4 可以得出标准的 PSO 算法无效进化代数多,且易陷入局部最优,无法收敛到全局最优解,而改进的 PSO 算法收敛的最优解略高于标准的 PSO,在不到 600 代就达到了最优解。

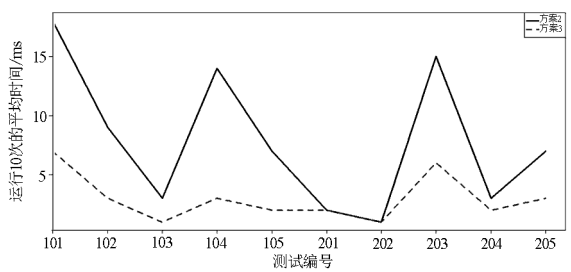

图 3 通过运行时间比较 3 种方案的匹配效率

Fig.3 Comparison of matching efficiency under the running time among three schemes

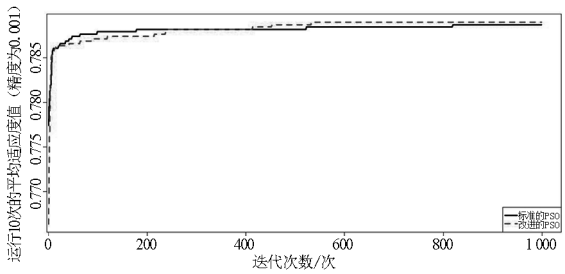

图 4 基于标准 PSO 和改进 PSO 的适应度值对比图

Fig.4 The comparison of the fitness value between standard PSO and improved PSO

综上所述,本文提出的两个创新技术(利用句法结构过滤案例候选集和利用改进的 PSO 优化 KNN 参数)不仅能够确定的质量优于传统 KNN 和基于传统 PSO 的 KNN 方案的案例匹配结果,还能显著提高案例匹配过程的效率。因此,改进 PSO 的 KNN 检索策略在案例推理中能更高效率的获取到高质量的案例匹配结果。

4 结论

针对 KNN 检索策略中检索效率低、缺乏隐含语义以及有人为因素的权重取值 3 个缺陷,提出了一种基于本体和微粒群算法的改进 KNN 检索策略。首先,基于图书馆领域本体的背景条件下,利用经过中文分词及查询扩展步骤得到的检索词集合中的谓语过滤案例库,得到初步案例候选集;然后为了避免相似度度量技术中参数设定包含人为因素影响,提出并设计了一种改进的微粒群优化算法,并利用改进 PSO 优化度量技术中的参数以提高案例匹配的准确率。实验数据采用的是福州晓锋科技信息咨询有限公司提供的图书馆参考咨询测试数据。实验结果表明,相比于传统 KNN 和基于传统 PSO 的 KNN 方案,本方案有效地提高了案例匹配结果的查全率和查准率。

参考文献:

[1] 杨斌宇.基于案例的推理在智能决策支持系统中的应用[D].长春:吉林大学,2004.
[2] 王津津.案例推理在决策支持系统中的应用研究[D].合肥:合肥工业大学,2010.
[3] 李锋刚,倪志伟,郜峦.基于案例推理和多策略相似性检索的中医处方自动生成[J]. 计算机应用研究,2010,27(2):544-547.
[4] 张春晓.案例推理的认知改进策略及学习性能研究[D].北京:北京工业大学,2014.
[5] 李小展.基于文本挖掘的医学诊疗案例推理系统的研究与应用[D].广州:广东工业大学,2011.
[6] 费玉莲.面向电子商务的谈判支持系统研究[D].杭州:浙江工商大学,2011.

[7] 李景.领域本体的构建方法与应用研究[D].北京:中国农业科学院,2009.
[8] 崔巍.基于 Peer-to-Peer 网和地理 ontology 的系统集成和互操作研究[J].计算机工程与应用,2003,39(32):45-47.
[9] 薛醒思.基于进化算法的本体匹配问题研究[D].西安:西安电子科技大学,2014.
[10] 杨美荣,邵洪雨,史建锋,等.改进的领域本体概念相似度计算模型研究[J].情报科学,2014,32(5):72-77.
[11] 唐中林.基于本体的概念相似度计算方法的研究[D].武汉:武汉理工大学,2013.
[12] 陈沈焰,吴军华.基于本体的概念语义相似度计算及其应用[J].微电子学与计算机,2008(12):96-99.
[13] 董颖,唐加福,许宝栋,等.一种求解非线性规划问题的混合粒子群优化算法[J].东北大学学报(自然科学版),2003,24(12):1141-1144.
[14] 关圣涛,楚纪正,邵帅.粒子群优化算法在非线性模型预测控制中的研究应用[J]. 北京化工大学学报(自然科学版),2007,34(6):653-656.
[15] 王书斌,单胜男,罗雄麟.基于 T-S 模糊模型与粒子群优化的非线性预测控制[J]. 化工学报,2012,63(S0):176-187.
[16] Shi Yuhui, Eberhart R C. Fuzzy adaptive particle swarm optimization[C]//Proceedings of the 2001 Congress on Evolutionary Computation. Washington D C: IEEE,2001:101-106.
[17] 刘伟,周育人.一种改进惯性权重的 PSO 算法[J].计算机工程与应用,2009,45(7):46-48.
[18] 申丹丹,石跃祥,周文杰,等.基于适应值引导的粒子群改进算法[J].计算机工程与应用,2015,51(14):63-66.
[19] 龚小谨,罗振声,骆卫华.汉语句子谓语中心词的自动识别[J].中文信息学报,2003,17(2):7-13.
[20] 张帆,钟金宏,黄玲.改进的领域本体概念相似度计算方法[J].计算机工程,2010,36(23):66-68.
[21] 胡建秀,曾建潮.微粒群算法中惯性权重的调整策略[J].计算机工程,2007,33(11):193-195.
[22] 黄少荣.粒子群优化算法综述[J].计算机工程与设计,2009,30(8):1977-1980.
[23] 王杰文,李赫男.粒子群优化算法综述[J].现代计算机(专业版),2009,30(2):22-27.
[24] 王家超.基于事例推理在甲型 H1N1 流感诊断中的应用研究[D].沈阳:东北大学,2010.

(特约编辑:黄家瑜)