第34卷 第24期 计 算 机 工 程 2008年12月
Vol.34 No.24 Computer Engineering December 2008
·博士论文· 文章编号：1000—3428(2008)24—0028—03 文献标识码：A 中图分类号：TP182

基于偏好信息的案例检索算法

李 锋[1]，魏 莹[2]
(1. 华南理工大学工商管理学院，广州 510640；2. 香港中文大学系统工程与工程管理系)

摘 要：案例推理方法建立在“相似问题具有相似解”的基础上，能否从案例库中检索出与新问题“最相似”的案例是案例推理方法成功的关键因素之一。该文提出一种改进的检索方法，在原始最近相邻算法基础上，用专家对新问题案例与历史案例属性差异的效用评价替代原始的属性差异值来衡量专家对属性差异的敏感程度。引入变异系数来标度新问题案例与历史案例的属性差异的分布情况，从而保证检索出的最相似案例具有较高的属性差异的均衡性。通过具体案例检索实例分析，验证了该方法的有效性。

关键词：案例检索；偏好信息；案例推理

Case Retrieval Algorithm Based on Preference Information

LI Feng[1], WEI Ying[2]
(1. School of Business Administration, South China University of Technology, Guangzhou 510640;
2. Department of Systems Engineering and Engineering Management, Chinese University of Hong Kong)

【Abstract】 Case-Based Reasoning(CBR) is a problem solving technique that solves new problems by finding a similar past case, and reusing it in the new problem situation. Noticing the shortcomings of K-Nearest Neighbors algorithm, this paper proposes an improved case retrieval algorithm measuring the similarity of new problem case and historical cases of case-based not only by difference of their feature values, but also by attribute(sensitivity) of reasoner. In this improved algorithm, original distance of feature of two cases is replaced by the reasoner's utility value of the distance value. In addition, a variance indicator is introduced to balance equilibrium degree of all differences. The proposed algorithm is compared with other popular algorithms and verified numerically.

【Key words】 case retrieval; preference information; Case-Based Reasoning(CBR)

案例推理(Case-Based Reasoning, CBR)方法建立在“相似问题具有相似解”的假设上，是通过调整或重用历史问题的解决方案来解决新问题的一种推理方法。在案例推理的“4R”问题求解流程中，首先是从案例库的诸多案例中，按新问题案例的描述特征检索(Retrieval)出与新问题最相似的历史案例。随后通过重用(Reuse)和修正(Revise)2个环节调整相似案例的解使之适用于新问题。当新问题的解决方案被确认有效后，新问题和适用解被保存(Retain)在案例库中，以备将来使用。

1 研究现状

案例检索的核心是如何针对具体领域定义和量化新问题案例与历史案例之间的“相似程度”。根据案例的表达和结构以及案例库的索引机制等因素的不同，案例检索有许多不同的技术选择，如最近相邻算法、归纳法、人工神经网络法、遗传算法等[1]。其中，最近相邻算法是应用最广泛的案例检索算法，但也有它的不足之处，例如：案例属性权重的取值没有统一标准；检索时间随案例数量呈线性增长；没有属性相似的具体计算公式。为此，结合具体应用领域和实际问题，许多专家学者提出了各有特色的改进方法。文献[2]提出的属性相似度算法考虑了案例库中该属性数值的分布情况，文献[3]则考虑了分布式环境下异构案例表达下的案例检索算法。

但是，作为一种多属性决策算法，最近相邻算法及其现有的改进方法并没有充分考虑案例推理应用系统的使用者的主观偏好。虽然文献[4]中提到了归一化效用函数，但其本质是一种案例属性数据的归一化预处理方法。因此，本文提出基于偏好信息的案例检索算法，旨在提高检索算法的质量和检索结果的可信度。

2 基于偏好信息的案例检索算法

在决策理论中，主观偏好用以衡量方案属性值对特定决策者的实际价值，即属性的效用值。用效用值代替原始数据参与决策主要是考虑到以下2点因素：属性数据的边际效用递减规律；不同的决策者有着不同的属性效用变化曲线。而案例检索实际上也属于一种多属性决策问题：从历史案例库中检索出与新问题案例差异最小的案例。因此，在案例检索算法中引入偏好信息(效用理论)能够更真实地反映出案例推理应用系统的使用者对新问题案例与历史案例属性差异的敏感程度，检索出的“最相似”案例具有更高的可信度，易于被接受和理解。

基于偏好信息的案例检索算法通过计算新问题案例和案例库中历史案例的相似度，输出其中相似度计算值最高的一个/多个案例作为案例检索的结果。新问题案例与历史案例相似程度的计算公式为

基金项目：广州市哲学社会科学发展“十一五”规划课题基金资助项目(08B12)

作者简介：李 锋(1975－)，男，讲师、博士，主研方向：决策支持，建模与仿真，供应链管理；魏 莹，博士后

收稿日期：2008-05-20 **E-mail：**fenglee@scut.edu.cn

$$S_{(X,Y)}=(1-\alpha)\cdot\sum_{i=1}^{n}\left(\omega_i\cdot u_i\left(d_i\left(f_i^X,f_i^Y\right)\right)\right) \tag{1}$$

其中，f_i^X和f_i^Y分别表示新问题案例X和历史案例Y的第i个属性的属性值；$d_i()$为新问题案例与历史案例在第i个属性值上的距离计算公式；$u_i()$为案例推理应用系统的使用者对于新问题案例与历史案例在第i个属性值上的距离$d_i()$的效用函数；ω_i表示案例第i个属性在整个案例属性集合中所占的权重，且$\Sigma\omega_i=1$；α为变异指标系数。

新问题案例与历史案例在第i个属性值上的距离计算公式$d_i()$属性值的特点如下：

精确值与精确值之间的距离计算公式为

$$d(a,b)=|a-b| \tag{2}$$

精确值与区间值之间的距离计算公式为

$$d'\left(a,[b_1,b_2]\right)=\left(\int_{b_1}^{b_2}d(a,x)\mathrm{d}x\right)\Big/(b_2-b_1)=\begin{cases}(b_2+b_1-2a)/2 & a\le b_1\\ \dfrac{(b_2-a)^2+(b_1-a)^2}{2\cdot(b_2-b_1)} & b_2>a>b_1\\ (2a-b_2+b_1)/2 & a\ge b_2\end{cases} \tag{3}$$

式(3)即精确值a与区间$[b_1,b_2]$内所有精确值距离的算术平均。

区间值与区间值之间的距离计算公式为

$$d''\left([a_1,a_2],[b_1,b_2]\right)=\frac{\int_{a_1}^{a_2}d'\left(x,[b_1,b_2]\right)\mathrm{d}x}{(a_2-a_1)\cdot(b_2-b_1)} \tag{4}$$

式(4)即区间与区间所有精确值距离的算术平均。

不失一般性，假定$a_1<b_2$，得到区间值与区间值之间的距离计算公式为

$$d''\left([a_1,a_2],[b_1,b_2]\right)=\begin{cases}(b_2+b_1-a_1-a_2)/2 & a_1<a_2\le b_1<b_2\\ \dfrac{b_2+b_1-a_1-a_2}{2\cdot(b_2-b_1)\cdot(a_2-a_1)} & a_1<b_1<a_2\le b_2\\ \dfrac{2\cdot a_1\cdot a_2-(b_2+b_1)\cdot(a_2+a_1)+b_1^2+b_2^2}{2\cdot(b_2-b_1)} & b_1<a_1<a_2<b_2\end{cases} \tag{5}$$

当$a_1\ge b_2$时，式(5)中$[a_1,a_2]$与$[b_1,b_2]$互换位置即可。

定义案例属性的效用函数$u_i()$取值范围为[0,1]：当新问题属性与历史问题属性差异/距离$d_i()$为 0 时，其效用值为 1；当$d_i()$无穷大时，其效用值为 0；其他情况下，$d_i()$介于 0~1 之间，具体效用值可以通过提问的方法与用户确定。

案例属性的权重系数ω_i的确定也是案例检索算法的一个重要环节。不同的案例推理应用系统的使用者对于案例的各属性有着不同的偏好次序，因此，案例检索之前，首先由使用者采用 1~9 比率标度定义两两属性之间的相对重要性，然后构造案例属性的互反判断矩阵。只有通过一致性检验的互反判断矩阵才能计算生成属性的权重系数；否则，需要使用者调整属性之间的相对重要性标度，直到互反判断矩阵通过一致性检验，即随机一致性比率参数小于 0.1。详细计算方法可以参见层次分析法中对判断矩阵的介绍。

最近相邻算法通过加权和的形式将多属性决策问题转化为一个单一属性的方案排序问题，隐含了目标属性的属性值之间的可补偿性，而且这种补偿是线性的。而实际上，许多决策问题的属性值之间是不可补偿的，即使在一定范围内可以补偿，也是非线性的。例如：对于消费者来说，一件质料优秀但样式很差的衣服的吸引力不会强于一件质料良好、样式也良好的衣服。即质料的优秀并不能弥补样式上的缺憾。

因此，本文引入统计学中综合反映现象总体各单位之标志差异程度的指标——变异指标来量化属性之间的不可补偿性。从而使得检索出来的相似案例的各个属性效用值具有较高的均衡性。变异指标的计算公式为

$$\alpha=\sigma_{(X,Y)}\Big/\bar{X}_{(X,Y)}$$

$$\bar{X}_{(X,Y)}=\sum_{i=1}^{n}u_i\left(d_i\left(f_i^X,f_i^Y\right)\right)\Big/n$$

$$\sigma_{(X,Y)}=\frac{\sqrt{\sum_{i=1}^{n}\omega_i\cdot\left(u_i\left(d_i\left(f_i^X,f_i^Y\right)\right)-\bar{X}_{(X,Y)}\right)^2}}{n} \tag{6}$$

其中，$\bar{X}_{(X,Y)}$为新问题案例与历史案例属性距离的效用值的平均值；$\sigma_{(X,Y)}$为加权形式下的新问题案例与历史案例属性距离的效用值的均方差。

3 实例

本文采用文献[4]所列的案例以及案例属性权重。同时，假定用户对于案例各属性差异的效用函数为指数函数：

$$u_i\left(d_i\left(f_i^X,f_i^Y\right)\right)=\mathrm{e}^{-d_i\left(f_i^X,f_i^Y\right)/\bar{d}_i}$$

$$\bar{d}_i=\sum_{j=1}^{m}d_i\left(f_i^X,f_i^Y\right)\Big/m \tag{7}$$

其中，$\bar{d}_i$表示案例库中所有历史案例(m 个案例)第 i 个属性与新问题案例的 i 个属性距离的平均值。

从指数函数曲线可以看出，指数函数型效用曲线表明用户对于新问题案例与历史案例属性差异比较敏感。

表 1 列出了案例库(编号为 1~18 的案例)与目标案例(编号为 19~23 的案例)最相似的 3 个案例相似度和文献[4]的计算结果。

表 1　检索结果

新问题案例		第 1 相似案例	第 2 相似案例	第 3 相似案例
19	案例编号	4	5	16
	相似度	0.748 5	0.638 9	0.618 2
	文献[4]	4	16	3
20	案例编号	15	10	14
	相似度	0.646 6	0.643 2	0.508 0
	文献[4]	15	10	7
21	案例编号	5	14	15
	相似度	0.568 8	0.526 3	0.510 8
	文献[4]	7	9	14
22	案例编号	6	16	5
	相似度	0.634 5	0.577 5	0.560 9
	文献[4]	6	5	16
23	案例编号	11	15	10
	相似度	0.803 3	0.454 0	0.440 4
	文献[4]	11	10	15

文献[4]中所采用数据归一化处理的 S 型曲线有如下特点：当新问题案例与历史案例属性的值远离历史案例中该属性的平均值，且处于该平均值的同侧时，其属性差值被该函数缩小；而当新问题案例与历史案例属性的值分处于平均值的两侧，或一个值靠近平均值而另一个值远离平均值时，其属性差值被函数放大。考虑到该归一化处理函数的这一特性，认为对于与编号为 19, 20, 22, 23 的新案例最相似的 3 个历史案例来说，本文所采用的算法检索出的结果与文献[4]得出的结果类似，可以接受。而对于与案例 21 相似的历史案例，本方法给出的结果与文献[4]给出的结果却有较大的差异：本文给出的最相似案例为案例 5，而文献[4]给出的是案例 7。

编号为 21, 5, 7 的案例的属性数据以及参与相似度计算

的各属性的权重见表 2。表中也列出了案例 21 和案例 5、案例 21 和案例 7 的原始属性差异。

表 2　案例数据

编号		属性 1	属性 2	属性 3	属性 4	属性 5
21	值	2.6	2.800	190	12.0	9.5
5	值	2.0	1.735	192	11.9	8.9
	差值	0.6	1.065	2	0.1	0.6
7	值	2.8	2.100	233	9.7	10.9
	差值	0.2	0.700	43	2.3	1.4
权重		0.291 6	0.149 9	0.103 0	0.112 7	0.342 8

从表 2 的数据来看，案例 5 有 3 个属性明显优于案例 7，而且一个属性略差于案例 7。因此，认为案例 5 与案例 21 的相似度高于案例 7 与案例 21 的相似度。即本文的检索结果在案例 21 上要优于文献[4]给出的结果。同时，表 3 给出了变异指标对案例检索结果的影响，表中“无”代表没有加入变异指标的与新问题案例最相似案例的相似度排名，“有”代表加入变异指标后的案例相似度排名，即表 1 中的排名顺序。

表 3　变异指标的影响

新问题案例		第 1 相似案例	第 2 相似案例	第 3 相似案例	第 4 相似案例
19	有	4	5	16	17
	无	4	16	17	5
20	有	15	10	14	7
	无	10	15	7	14
21	有	5	14	15	12
	无	5	12	15	14
22	有	6	16	5	4
	无	6	16	5	4
23	有	11	15	10	14
	无	11	15	10	14

以与新问题案例 20 最相似的 2 个案例为例，加入变异指标后，最相似案例为 10，而之前为案例 15。表 4 给出了经过指数效用函数处理后的案例 10、案例 15 与案例 20 各属性差异效用值分布。

表 4　属性效用差异表

编号	属性 1	属性 2	属性 3	属性 4	属性 5
10	0.605 4	0.443 2	0.722 6	0.765 9	0.314 6
15	0.776 1	0.720 6	0.822 9	0.548 8	0.259 5
权重	0.291 6	0.149 9	0.103 0	0.112 7	0.342 8

结合表 5 中给出的案例 20、案例 10、案例 15 的属性数据和差值，认为案例 15 与案例 20 具有更高的相似程度。同时此结果也与文献[4]中给出的相似结果一致。因此，变异指标的引入，增加了本文所介绍的案例检索算法结果的可信性。

表 5　案例数据

编号		属性 1	属性 2	属性 3	属性 4	属性 5
20	值	3.000	2.500	200	11.0	12.0
15	值	2.800	2.500	193	12.9	13.0
	差值	0.200	0.000	7	1.9	1.0
10	值	2.996	2.055	185	11.6	12.5
	差值	0.004	0.445	15	0.6	0.5
权重		0.291 6	0.149 9	0.103 0	0.112 7	0.342 8

4　结束语

本文介绍了一种案例检索算法，用新问题案例与历史案例属性的距离的效用值替代原始的距离参与相似度计算，能够更好地体现使用者对属性差异的敏感程度。同时，该算法还考虑了使用者对各属性的距离的均衡性要求。具体实例的计算结果表明了算法的有效性和可信度。

参考文献

[1] 李　锋, 冯　珊. 基于人工神经网络的案例检索与案例维护[J]. 系统工程与电子技术, 2004, 26(2): 234-236.

[2] 李　锋, 周凯波, 冯　珊. 基于统计特征的属性相似度计算模型[J]. 华中科技大学学报: 自然科学版, 2005, 33(6): 80-82.

[3] 李　锋, 魏　莹. 分布式环境下基于语义相似的案例检索[J]. 计算机工程, 2007, 33(9): 28-30.

[4] 罗忠良, 王克运, 康仁科等. 基于案例推理系统中案例检索算法的探索[J]. 计算机工程与应用, 2005, 41(25): 230-232.

(上接第 27 页)

子系统的状态转换如图 6 所示。

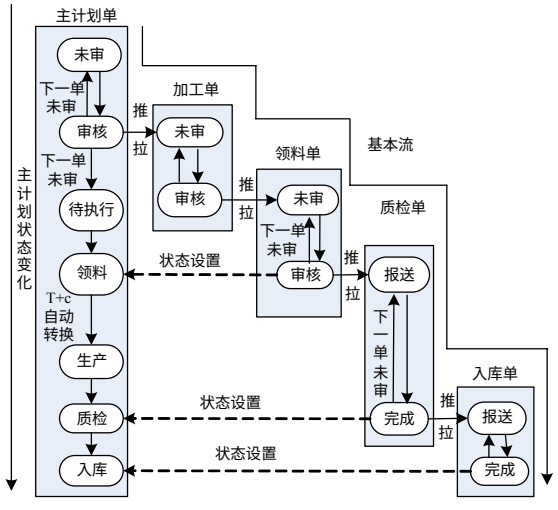

图 6　生产子系统状态转换

5　结束语

本文在研究与实践的基础上，设计开发了基于.NET 的面向中小制造业的 SmartERP 系统。系统具有多个可重用组件，包括数据持久层、虚拟数据访问服务和显示工具集等，提供多种实用服务，包括对象缓存、国际化等，并预留了与电子商务系统结合的接口，能迅速针对特定企业进行二次开发。

参考文献

[1] 周才堂. 我国汽车制造企业 ERP 需求分析与解决方案[EB/OL]. (2007-05-08). http://e.chinabyte.com/22/1806022.shtml.2004-06-08/2007-5-8.

[2] 胡嘉贤, 常会友, 衣　杨. 支持功能可重构的 ERP 开发平台中的构件技术[J]. 计算机应用, 2006, 26(7): 1738-1743.

[3] 朱战立, 王魁生. 中小企业 ERP 软件系统框架研究[J]. 计算机工程, 2006, 32(12): 37-38.

[4] 郑彦树. 基于构件的可重构 ERP 系统研究[J]. 计算机工程与设计, 2006, 27(17): 3168-3171.

[5] Terrylee. 设计模式[EB/OL]. [2007-05-10]. http://www.terrylee.cn.