文章编号:1674 - 2869(2010)03 - 0096 - 04

案例知识库维护技术的研究进展

李建洋[1],倪志伟[1],郑金彬[2],谢秀珍[2]
(1.合肥工业大学计算机网络所,安徽 合肥 230009;2.龙岩学院计算机科学系,福建 龙岩 364000)

摘　要:案例库是 CBR 系统的主要知识库,但是难以维护,一个主要因素是案例库大,并且是非结构化或半结构化的,用自然语言来表达的.文章针对案例知识库维护中出现的相关复杂问题,分析了各种维护策略的可行性,提出了在不同环境下实施维护的准则,指出了选用合适的方法来实现案例库维护.

关键词:案例库维护;不确定删除法;选择删除法;粒度计算
中图分类号:TP39;TP181　　文献标识码:A　　**doi**:10.3969/j.issn.1674-2869.2010.03.025

0　引　言

案例推理(Case-Based Reasoning, CBR)是由目标案例而得到历史的源案例,并由此来指导目标案例求解的一种策略;它从另一个侧面实现了人类智能,绕过了"知识获取"这个难题,因此克服了基于规则系统(Rule-Based Reasoning, RBR)的一些弱点,是一种重要的机器学习方法.它根据相似性原理,由一个已知系统具有某些属性,猜想另一个未全知系统也具有这些属性;使用类比推理模式和假设推理模式,是从特殊到特殊的推理过程.案例由诸多的属性组成,目标案例和源案例的本质特征具有相似性关系使得类比有了基础[1].

案例学习系统往往处理的是复杂领域的问题,案例的表示不具备高结构化和稳定性,CBR 系统的强大功能来源于它能从知识库中迅速检索出(case retrieve)相关案例[2-3].传统的数据库索引机制虽然可以有很好的借鉴作用,但却存在极大的差别:传统的数据库强调的是保持存储结构平衡,CBR 的索引用来在需要的情况下指出一个独立的案例,不再强调存储结构平衡,而关心如何把案例库划分成概念上有用的片段;传统的数据库的操作是精确匹配,CBR 中进行的却是相似性匹配[4-6].

1　维护技术

案例库作为 CBR 系统中的主要知识库,其学习功能即是不断往案例库中增加新的案例.当其不断增大时,带来的好处是很容易找出相同或相似案例,减少修正阶段(case adaptation)的次数与时间.一般来说知识库越大,知识越丰富,这样 CBR 系统可以解决更多的问题,体现了它的智能水平.但伴随着案例库的不断扩大,会导致相似案例的检索时间大大增加,引发"沼泽问题".案例库的维护就是指实现一些更新案例库组织结构或内容的策略,包括表达方式、领域内容、描述信息、实现方式,以保证未来的推理能完成特定的性能指标[7-9].

1.1　不确定删除法

1.1.1　随机删除法　当知识库中的规模超出一定的预先设定值时,就随机删除一个案例.随机地删除案例,可能是关键案例,从而导致系统能力的急剧下降,甚至有的目标问题永远无解.

1.1.2　实用值度量法　utility 效用＝平均节省时间 * 应用频率 － 匹配代价,根据效用值来决定是否删除.由于随着知识库的增加,一个特定知识的应用频率总是要下降的,所以它的度量值也在衰减,并且实用值较小的不一定就是无用的案例.

1.1.3　IB3 方法　该方法通过对案例库中的每个案例建立一个匹配记录,一旦记录指示该案例是一个无用案例,即把它从案例库中删去.此算法的缺点是:它是一个被动的不太精确的办法,它没有对每个案例的能力给予准确地评估.

1.2　选择删除法

1.2.1　基于案例分类的删除策略　该方法认为在案例库中并不是所有案例都个个平等,它通过计算一个案例的覆盖度(Coverage)及其可触及的程度(Reachability),区分出了 4 个类型的案例:核

收稿日期:2009 - 11 - 25
基金项目:863 专项课题(2007AA01Z116);国家自然基金(70871033);福建省科技专项课题(2008F5013);安徽省高校自然基金(KJ2008B107);福建省教育科研基金(JA08229)
作者简介:李建洋(1968 -),男,安徽合肥人,教授,博士,研究方向:机器学习、神经网络、智能系统.

心案例，连接案例，辅助案例，支持案例. 为控制案例库的规模，算法依次删除相对次要的案例.

1.2.2　基于模式归纳的案例库维护　在案例保存阶段(case retain)进行模式归纳，可以寻找类似案例的共性，再加以抽象和泛化. 通过模式归纳，从而可以在案例库中删除一些极为相似的案例. 在 CBR 系统在求解问题时，当有了抽象的一般化知识，可以不必借助于相似的具体案例，因而减少候选集合中源案例的数量.

1.2.3　维护规则方法、基于 Agent 等办法　Leake 和 Wilson 在 1998 年提出了在检索阶段使用维护规则来更新现有案例，以解决环境的快速变化而引起的案例库维护问题. Racine 和 Yang 在 1997 年描述了一个基于 Agent 的方法来检测冗余案例和冲突案例；他们又详细研究了半结构化的案例库上的维护问题，也采用了 Agent 技术. McSherry 叙述了一类利用案例知识获取工具 CaseMaker 来选择案例，它是从一组由大到小按案例覆盖度排序的案例中选择出的、并建立案例库的方法.

1.2.4　基于孤立点的案例库维护　孤立点就是对给定的数据对象进行分析，其中某些数据是显著相异的、异常的或不一致的，通常有基于统计、基于距离、基于偏离三种检测方法. 在案例知识库维护中，我们采用基于距离的孤立点方法，可以与基于相似距离的案例检索相一致，因而算法并不需要特别额外的时空开销[10-11]. 通过聚类等任何方法，剔除噪声或错误等确实无用的孤立点，保留了可靠的孤立点——非凡的案例知识，具有极高的泛化能力.

1.2.5　基于相似粗糙集技术的案例库维护　相似粗糙集技术是对粗糙集理论 RS(Rough Sets)的应用研究中，提出的一种扩展模型. 它不但继承了经典粗糙集的各种优势；同时以相似关系取代等价关系后，可以避免对案例属性数据的离散化处理(案例属性数值绝大多数是连续的，为此经典粗糙集采用了多种离散化方法，但是很容易造成数据的割裂)；而且它对相似关系的定义，与案例相似性的定义完全相同，可以和 CBR 系统完美地结合[12-13].

相似粗糙集技术可以有效地利用差别矩阵，通过不同的相似度阈值发现以及处理案例库的冗余，从而可以有选择地删除满足阈值的多余的相似案例，保证了案例库拥有较高的覆盖度；同时减少了案例库维护过程中相似度的计算量，并且可以实现案例库的动态维护.

1.3　非删除法

如上所述，不确定删除法虽然可以限制案例库的无限膨胀，但是效果并不可靠，因此使用受到限制. 选择删除法是基于这样的假设：随着案例系统的学习，会有各种冗余的案例加入案例库，因此可以使用某些策略来搜寻并将其永久地删除，只需保留一些符合某些标准的“高质量”的案例. 选择删除法是目前案例库维护的主要手段，但是删除法或多或少是以牺牲知识库为代价，以换取 CBR 系统推理时间和空间的平衡.

然而，在一些电子商务、网络 CBR 以及一些交互式 CBR、分布式 CBR 应用领域，诸如故障诊断、联机决策等具体的应用中，案例库中的每一个案例都代表一个唯一的商品、或者一个不可缺失的宝贵经验，案例库很容易达到成千上万的规模，而且都是不可约简的，显然上述案例库维护方法是不能应用在这些环境中.

1.3.1　交叉覆盖法　CBR 与神经网络之间具有天然的密切联系，人们已经研究开发了多种在 CBR 系统中应用神经网络的方法. 但是上述的应用系统存在着一些难以克服的弱点，如系统的可解释性较差、系统实现复杂、实用性较差等，特别是系统中因神经网络的算法复杂性较高而难以用于 CBR 知识丰富的大规模案例库系统. 交叉覆盖算法是构造性神经网络算法，采用多层前馈神经网络技术，解决了多年来一直未解决的作为分类器的多层前馈网络的设计问题，不但具有很高的分类识别率，而且时间与空间的复杂度低，因此可以作为大规模、高维数据量的分类器[14-15].

该方法首先通过扩维、空间投射方法，较好地实现案例库中的相似案例的领域覆盖，实现信息的选择性过滤. 其次，通过将这些获得的覆盖领域，输入到多层前馈神经网络中实现案例匹配，提高检索效率. 该方法并没有缩减案例库，通过使用易于构造、易于理解的多层前馈神经网络，并且采用交叉覆盖算法来有效地降低网络的算法复杂度，建立起了一种可信赖的高性能、确保案例的性能与效率，可以有效地解决因学习导致的案例库规模增长而产生的问题.

1.3.2　商空间法　知识的粒度性是造成使用已有知识不能精确地表示某些概念的原因[16-17]，目前粒计算的主要模型是模糊集、粗糙集和商空间模型. 与粗糙集类似，商空间理论也使用等价关系来描述，但其独到之处在于不只是研究二维的关系，还研究了对象之间的结构关系. 由于考虑到论域的结构，借助于拓扑中连通性以及映射的连续

性，可以得到该推理模型具有的最重要的性质——同态原则，即保真原理(或保假原理)[18]. 当面对一个复杂问题时，人们因而可以通过合理的分层递阶，大大降低问题求解的复杂性.

在案例系统应用的复杂的决策科学领域，人类对事物的认识属性的先后和关注度不同，研究基于商空间粒度变换的案例推理系统，实现推理知识的粒度变换，(1)可以获取局部知识决策，克服CBR智能系统中因知识库信息缺失、决策信息不完全带来的推理难题；(2)可以完成概略地、由粗到细、不断求精的多粒度分析法，避免了计算复杂度高的难题，从推理需要的不同知识层次来研究问题.

这样，应用商空间粒度模型所获取的典型意义就在于——可以实现类似结构化数据库的案例知识检索片断，方法是：建立基于不同粒度知识的类似决策树，直接实现粗到细的案例知识检索，配合复合知识的多粒度合成，极大地降低检索复杂度. 如前所述的交叉覆盖法就是一种运用知识粒度的检索，但该方法中的知识只是基于某种"定"粒度的划分，尚未运用不同粒度的变换；而商空间法可以提供动态的知识粒度的变换.

1.4 维护原则

随着应用的日益开展，许多实用的CBR系统，广泛应用于医疗诊断、电路或机械设计、故障诊断、气象等各个领域；大型的CBR系统也越来越普遍了. 因此CBR系统的案例库运行效率问题突出，而且由于噪音案例以及错误案例的存在，最终会导致系统总体的性能降低，因此案例库的能力以及效率是判断案例库维护质量的依据. 目前的研究实现了案例库维护在案例推理过程中的重要性，案例作为专家经验的计算机存储形式，是人类三种思维(直觉、逻辑、创造性思维)的一种综合表现形式.

由于存在于人类思维中的非单调逻辑推理，属于非标准逻辑范畴；用案例形式表示，容易发现冗余的和不一致知识；尽可能对此进行实时监测，实现实时维护. 在不断变化的环境中，由于领域知识的变化，导致类比的基础即由特殊到特殊的知识假设推理失效，只能选用删除法进行案例库维护. 针对实际领域应用中不可约简案例库的CBR系统性能维护难题，只有依赖两方面的突破：改进案例检索算法以适应大规模的案例库，对案例库采取某种过滤手段以适应案例检索算法，从而确保CBR系统的可靠、高效运转.

2 结 语

案例知识库是一个系统及组织的核心财富，也是CBR系统研究的核心难题，涉及CBR推理的知识表示、适配与改写过程；由于是非结构化的，难以通过常规方法实现维护. 案例知识库维护作为CBR研究的一个重要分支，已经开发出来不同的维护策略；然而不同的环境下，因CBR系统的规模、时效性以及应用领域的特点不同，其维护手段和维护性能存在较大的差异.

随着应用领域的拓宽和应用程度的加深、规模的不断增加，其维护手段日趋复杂. 目前的国内外研究中，特别是国际商业化的CBR应用系统，由于长期持久地运行，对CBM技术提出了种种挑战. 随着人工智能领域研究的不断进步，融合其它学习技术的案例推理系统(如机器学习、神经网络等)已经迅速展开. 未来的案例库维护技术应当可以全面地监控CBR的能力，如案例库的增长速度及系统的性能检查；案例库维护的定量化分析与证明；动态地组织与更新案例库及案例库实时压缩等技术的研究将会更加深入.

参考文献：

[1] 倪志伟. 智能管理技术与方法[M]. 北京：科学出版社，2007.

[2] Susan Craw, Jacek Jarmulak & Ray Rowe. Maintaining Retrieval Knowledge in a Case-Based Reasoning System[J]. Computational Intelligence, 2002,17(2): 346-363.

[3] Roth-Berghofer T, Reinartz T. MAMA: a maintenance manual for cased-based reasoning systems [C]// Proceedings of International Conference on Case-Based Reasoning 2001(ICCBR 2001). Springer, 2001:452-466.

[4] Petra Perner. Case-Based Reasoning and the Statistical Challenges[C]// Proceedings of European Conference on Case-Based Reasoning (ECCBR 2008). Springer, 2008:430-443.

[5] Grachten M,F Alejandro G, Josep L A. Navigating through case base competence[C]// Proceedings of International Conference on Case-Based Reasoning 2005(ICCBR 2005). Springer, 2005:282-295.

[6] Isabelle Bichindaritz. Prototypical Cases for Knowledge Maintenance in Biomedical CBR[C]// Proceedings of International Conference on Case-Based Reasoning 2007 (ICCBR 2007). Springer, 2007:492-506.

[7] Gomes P, Pereira F C, Paiva P. Evaluation of case-

based Maintenance Strategies in Software Design [C]// Proceedings of International Conference on Case-Based Reasoning 2003 (ICCBR 2003). Springer, 2003:186 - 200.

[8] Chen Fu-chien, Wen Chih-wang, Jen Chich-cheng. Data mining for yield enhancement in semiconductor manufacturing and an empirical study [J]. Expert Systems with Applications, 2007,33(1):192 - 198.

[9] Petra Perner. Case-base maintenance by conceptual clustering of graphs[J]. Engineering Applications of Artificial Intelligence, 2006,19(4): 381 - 393.

[10] Ni Zhi-wei, Liu Yu, Li Feng-gang. Case-base maintenance based on outlier data mining [C]// Proceedings of International Conference on Machine Learning and Cybernetics 2005 (ICMLC2005). Springer, 2005: 2861 - 2864.

[11] Shi Dong-hui, Zhang Chun-yang, Cai Qing-sheng. A new algorithm of outlier mining in data base [J]. Mini-micro Systems, 2002,22(10):1234 - 1236.

[12] 李建洋，倪志伟，刘慧婷. 一种基于相似粗糙集技术的案例库维护[J]. 计算机工程与应用，2005，41(32):19-21.

[13] Gento A M, Redondo A. Rough sets and maintenance in a production line [J]. Expert Systems, 2003, 20(5): 271 - 278.

[14] 张铃，张钹. 多层前馈网络的交叉覆盖设计算法[J]. 软件学报，1999,10(7):737 - 742.

[15] Li-Jian-yang, Ni Zhi-wei, LIU Xiao. Case - base maintenance based on multi - layer alternative - covering algorithm [C]// Proceedings of International Conference on Machine Learning and Cybernetics 2006 (ICMLC2006). Springer, 2006: 2035 - 2039.

[16] Zhang Ling, Zhang Bo, The structure analysis of fuzzy sets[J]. International Journal of Approximate Reasoning. 2005:24:92 - 108.

[17] Zadeh L A. Generalized theory of uncertainty (GTU)-principal concepts and ideas [J]. Computational Statistics & Data Analysis, 2006, 51(1): 15 - 46.

[18] 张铃，张钹. 模糊商空间理论(模糊粒度计算方法)[J]. 软件学报，2003,14(4):770 - 776.

Some advances in maintenance research of case knowledge base

***LI Jian-yang*[1], *NI Zhi-wei*[1], *ZHENG Jin-bin*[2], *XIE Xiu-zhen*[2]**

(1. Institute of Computer Network, Hefei University of Technology, Hefei 230009, China;
2. Department of Computer Science, Longyan University, Longyan 364000, China)

Abstract: With the dramatic proliferation of Case-Based Reasoning (CBR) systems in commercial applications, CBR has grown from a quite new area to a subject of major influence within Artificial Intelligence. Case bases are main knowledge sources in CBR systems, but they are difficult to maintain. One of the contributing factors is that these case bases are often large and yet unstructured or semi structured: they are represented in natural language text. Tackling the following problems related to the complexity of case base maintenance (CBM), the paper presents many kinds of strategies to achieve CBM and puts forward some criteria for this course, it suggests choosing the most suitable method for CBM based on its feasibility, which are under description in the face of different conditions.

Key words: CBM; indefinite deletion; selective deletion; granular computing

本文编辑：陈小平