

文章编号:1008-1542(2012)02-0150-04

## 分类算法在范例推理中的研究与应用

刘连喜<sup>1</sup>, 邢彤<sup>1</sup>, 徐浩<sup>2</sup>, 王伟<sup>3</sup>, 高凯<sup>3</sup>

(1. 河北科技大学财务处, 河北石家庄 050018; 2. 河北化工医药职业技术学院, 河北石家庄 050026; 3. 河北科技大学信息科学与工程学院, 河北石家庄 050018)

**摘要:**将范例推理中的范例初步匹配看作文本分类的特殊情形, 提出基于类别中心向量的分类算法。通过确定待处理案例的归属类别, 缩小范例检索范围, 减少在范例精确匹配阶段的计算量, 提高案例初步匹配的准确性。在此基础上, 将上述算法应用在对交通事故案例的处理与交通信息预警系统中。实验与使用表明, 该算法能较为准确地判断事故类型并给出相应的预警信息。

**关键词:**人工智能; 自然语言处理; 分类; 信息检索; 范例推理

中图分类号: TP274 文献标志码: A

## Study on category algorithm and its application in case-based reasoning

LIU Lian-xi<sup>1</sup>, XING Tong<sup>1</sup>, XU Hao<sup>2</sup>, WANG Wei<sup>3</sup>, GAO Kai<sup>3</sup>

(1. Finance Department, Hebei University of Science and Technology, Shijiazhuang Hebei 050018, China; 2. Hebei Chemical and Pharmaceutical College, Shijiazhuang Hebei 050026, China; 3. College of Information Science and Engineering, Hebei University of Science and Technology, Shijiazhuang Hebei 050018, China)

**Abstract:** This paper presents the class-centre vector based category algorithm, which is regarded as the basis of the text classification in case-based reasoning application. On the basis of the proposed algorithm, this method can be used to determine those cases' affiliation, and as a result this can reduce the retrieval scope so as to reduce the calculation and improve the precision. We use the proposed algorithm in the traffic accident analysis and the corresponding early warning system. The experimental results and the analysis prove the feasibility of the approach, and the existing problems and further works are also discussed in the end.

**Key words:** artificial intelligence; natural language understanding; category; information retrieval; case-based reasoning

范例推理(case-based reasoning, CBR)作为基于规则推理技术的一个重要补充,是当前人工智能及机器学习领域中的热门课题与前沿研究方向之一。CBR的执行过程如图1所示,主要包括检索相似案例、重用相似案例并推断新案例解决方案、修订解决方案、保存新案例以备后用等。由于CBR具有信息的完全表达、增量式学习、形象思维的准确模拟、求解效率高等特点,因此能够应用在那些没有很强的理论模型和领域知识不完全的决策环境中<sup>[1]</sup>。目前,关于范例推理的研究主要集中在以下几个方面:范例的索引及检索技术,范例修正技术及其修正规则的获取方法,范例库的维护技术及其性能的研究,范例工程的自动化,范例推理的理论基础,范例推理与其他方法(包括学习技术、多Agent技术、推理方法、数据挖掘、数据仓库技术)的集

收稿日期:2011-11-17;责任编辑:李穆

基金项目:河北省科技支撑计划资助项目(074572172)

作者简介:刘连喜(1965-),女,河北高邑人,高级会计师,主要从事信息管理、会计理论与电算化技术等方面的研究。

通讯作者:高凯副教授。E-mail:gaokao68@163.com

成技术,范例推理的应用,研制 CBR 开发平台, CBR 融合及大规模并行处理,基于 Web 的分布式 CBR 系统,可视化 CBR 技术及对话式 CBR 模型等<sup>[2-5]</sup>。

由于交通事故发生的原因和案发现场的勘察结果之间往往有着一定的因果联系,因此通过对事故勘察的特征提取,并通过相应的匹配算法,将所抽取出来的新案例特征与范例库中典型事故案例——指那些已成功处理、具有某些典型的特征属性进行比较,检索那些较符合的事故案例,找出相应事故的解决方案,并由分析得出相应的预警信息是可行、可信的。基于 CBR 的交通事故处理及其预警机制研究,可为处理类似的交通事故案件提供决策参考。但一般的范例推理过程在检索源范例时,往往存在着检索结果不准确、检索效率低下等问题。如果能在检索范例之前,通过适当的分类算法,确定新案例最可能归属的类别,然后再在其所属类别范围里通过相似度查找最近似的部分范例,就可以有效缩小范例检索的范围,提高范例检索的效率和准确率。在此将范例的初步匹配问题看作文本分类问题的特殊情形,提出基于类别中心向量的分类算法,并将其应用到范例推理的初步匹配阶段。从空间几何角度看,计算平均向量的过程就是计算多维空间中 1 个点集群中心的过程,只要 1 个点处于以这个中心和一定半径所划定的空间区域里,就可以认为该点所代表的案例归属于这个中心向量所代表的类别。在前期工作基础上,通过分类算法首先确定待处理案例的归属类别,减少了在范例精确匹配阶段的计算量,提高了范例推理系统整体性能。通过对公路车辆监控与交通事故处理(数据分析)的应用表明,该算法能较为准确地判断事故类型并给出相应的预警信息。

## 1 范例的组织及其向量化表示

在 CBR 及其推理过程中,首先要明确范例库及其组织方式。在本应用中,范例库中所有交通事故范例按事故原因被分成 8 个类存储,相同类别的范例保存在以类别名称命名的文件夹中,各种信息用不同字段标志符表示出来。当有新的案例到来时,首先依据类别中心向量分类算法确定新范例的归属类别,然后计算新范例与其归属类别中每个范例的语义相似度,如果相似度大于阈值,就认为新范例与已有范例相似,不再储存,否则将新案例的信息数据以文本文件的形式保存到所属类别的文件夹里。

当完成了范例的向量化表示后,就构建起了范例库的向量空间模型,之后才能在这个向量空间里实现分类算法。为了便于将范例内容映射入向量空间,使用中科分词模块 ICTCLAS 对范例的正文文本进行分词处理。即首先将范例正文中的所有非中文字符替换为空格符,使用分词模块对预处理的范例正文进行分词,根据停用词词典去除分词结果中的停用词。要想实现范例的向量化表示,须首先建立向量模板,之后可以依据向量模板实现范例的向量化。向量模板的构建过程大致分如下 4 个步骤。

1) 逐篇扫描范例库每个范例正文的分词结果,分别统计各词项在所属范例正文中出现的词频(TF)和各词项在所属范例类别中出现的文档数(DF)。

2) 选择特征词。采用 TFIDF 算法作为特征词权值的计算方法,计算公式如式(1)所示,其中 IDF 是词条  $t$  的逆文档频率,  $m$  是某一类别文档中包含词条  $t$  的文档数,  $n$  是包含词条  $t$  的文档总数。

$$IDF = \log\left(\frac{m}{n} \times N\right). \quad (1)$$

3) 将所有词项按照 IDF 值从大到小排列,再根据预先设定的最大向量维数  $k$  选取 IDF 最大的前  $k$  个词项作为特征词项。为了统一所有特征词项的分布概率,需要对特征词的 IDF 值进行归一化处理,具体方法如式(2)所示,式中  $IDF_i$  表示第  $i$  项特征词的 IDF 值,  $\max(\cdot)$  函数表示所有特征词项中最大的 IDF 值。

$$IDF_i := \frac{IDF_i}{\max(IDF_i)}. \quad (2)$$

在构建了向量模板以后,就可以依据向量模板实现范例的向量化表示了,主要步骤如下。

1) 将范例正文中的非中文字符替换为空格,用分词模块对其进行分词。

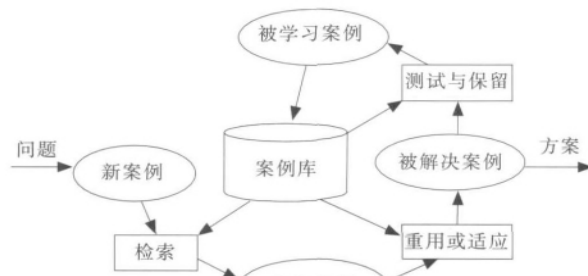


图 1 CBR 执行过程

Fig. 1 Process of CBR

2) 统计每个范例中各词项的文档频率 TF。考虑到范例正文长度的差异,须依据范例正文长度对每个词项的文档频率进行归一化处理,计算公式如式(3)所示,其中  $TF_i$  是范例中词项  $i$  的文档频率,  $l_i$  是范例正文长度。

$$TF_i := \frac{TF_i}{l_i} \tag{3}$$

为了统一范例中词项的分布概率,还需要对词项的 TF 值进行归一化处理,见式(4),式中  $\max(\cdot)$  是范例正文所有词项中最大的 TF 值:

$$TF_i := \frac{TF_i}{\max(TF_i)} \tag{4}$$

在完成向量归一化后,扫描范例分词结果中的每个词项。当把一个范例正文分词结果中所有词项扫描完毕后,就得到了该范例的空间向量。

## 2 类别中心向量的分类算法及其应用

类别中心向量是用各个类别的中心向量来代表相应类别,并且根据新来的案例向量与各类别中心向量的相似程度来判别新案例所归属的范例类别。算法处理流程见图 2。

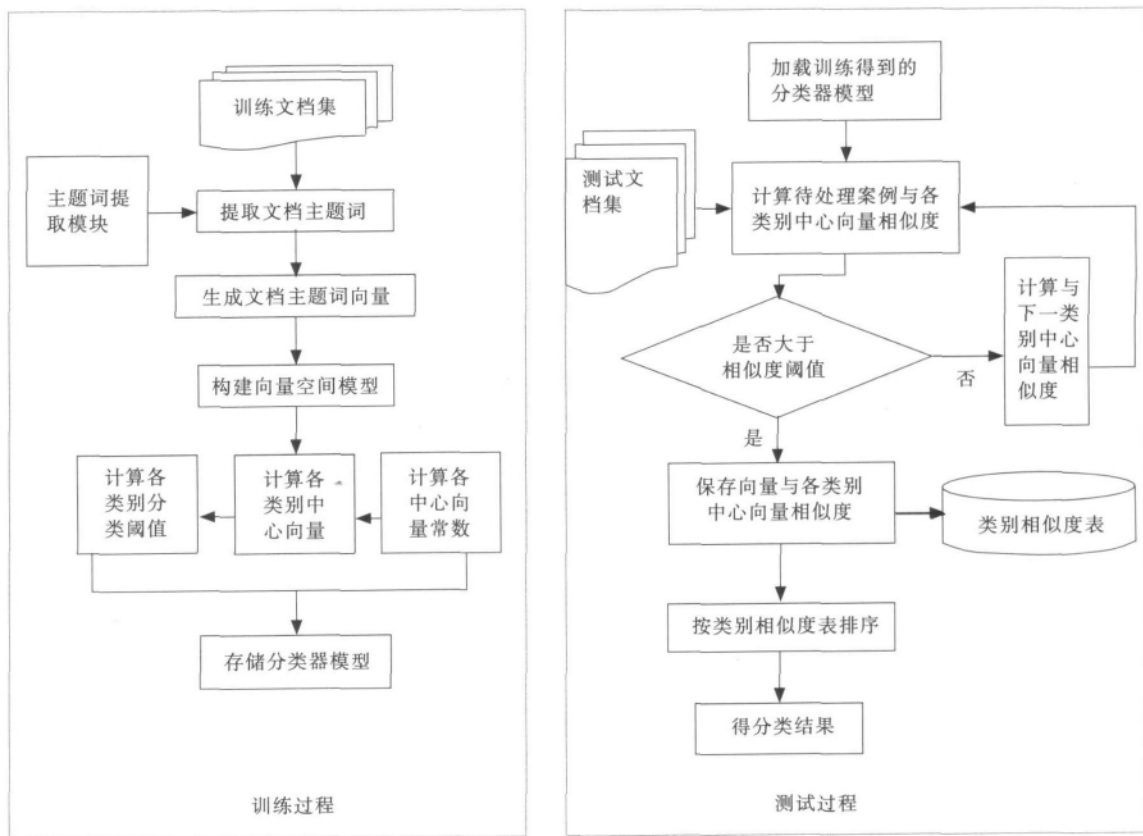


图 2 类别中心向量分类算法流程图

Fig. 2 Flow chart of the class-center vector based category algorithm

在建立了范例的向量空间模型以后,各范例都表示为向量的形式。笔者采用算数平均的方法来计算各类别的中心向量,计算公式如式(5)所示,式中  $V_{center}$  为某文档类别的中心向量,  $W_1$  为范例向量中第 1 维向量的权重,  $W_j$  为范例向量中第  $j$  维向量的权重,  $n_i$  为文档类别  $i$  中包含的文档数:

$$V_{center} = \left( \frac{\sum_{i=1}^n W_1}{n_i}, \dots, \frac{\sum_{i=1}^n W_j}{n_i}, \dots, \frac{\sum_{i=1}^n W_m}{n_i} \right) \tag{5}$$

通过上述步骤计算得到了各范例类别的中心向量、中心向量常数以及各类别的分类阈值后,就可以着手生成类别中心向量分类算法中的分类器模型。为了便于各种类型数据的统一存储,采用 XML 文件格式保存分类器模型。当某个类别的中心向量需要动态修正时,XML 文件可以方便地读取和保存相应类别的中心向量。图 3 是分类流程。范例推理系统的输入数据就是待处理的交通事故案例,输出就是交通事故的处理方案和某一时段的交通预警信息。在得到(或者用户输入)新案例后,首先将新案例的数据进行向量化处理,然后应用提出的分类算法来确定新案例所属的类别,从而实现案例的初步匹配。之后,就可以在新案例所属的类别内利用向量的相似度算法为新案例找到最为相似的范例,从而实现案例的匹配,最终实现交通事故案例的范例推理过程,并借助它辅助交通处理事故。

### 3 实验分析与测试

分别在不同的训练和测试样本集中,在不同的兼取类别情况下,对类别中心向量分类算法在交通范例推理系统中的应用性能进行了测试。表 1 是在封闭性测试情况下,测试兼取类别数对分类器性能的影响。

表 1 封闭性不同兼取类别测试

Tab. 1 Colse test on different compatibility numbers of the category

测试类别	兼取类别数	修正中心向量	总体准确率 $P$	总体查全率 $R$	总体 $F_1$
封闭性测试	1	修正	0.828 0	1	0.906 0
封闭性测试	3	修正	0.996 7	1	0.998 3
性能提高			20.37%	—	10.19%

从表 1 数据可以看出,兼取类别数增加后,分类器的性能得到了显著提高。可见在对交通案例进行初次匹配时,必须考虑兼类的情况,这也是符合客观事实的。

在交通案例的选取上,基础案例的缺乏是一个不可避免的问题。由于行业数据的保密性,这里采用的是从网上收集的一些案例来充当范例库的素材。尽管本文提出的算法有效实现了案例的初步匹配,但是该分类算法的性能还有提升空间,所以下一步还要在改进分类算法性能上再做一些工作。

### 4 结 语

将范例的初步匹配问题看作文本的分类问题,提出基于类别中心向量的分类算法,并将其应用到范例推理系统的初步匹配阶段。通过分类算法首先确定待处理案例的归属类别,有效缩小了范例检索的范围,减少了在范例精确匹配阶段的计算量,提高了范例推理系统整体性能。实验与使用表明,该算法能较为准确地判断事故类型并给出相应的预警信息。

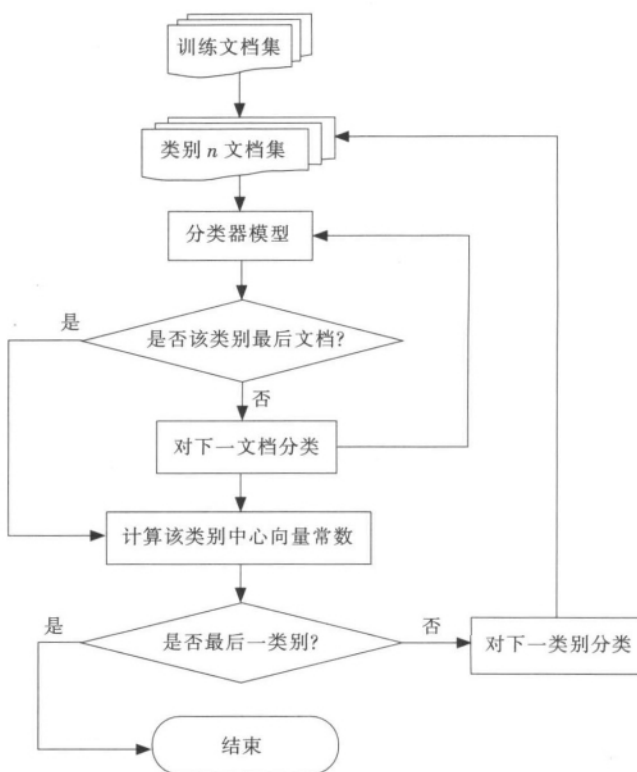


图 3 分类流程

Fig. 3 Flow of the category

(下转第 183 页)

温度越高,解吸时间越长,对  $\text{SO}_2$  的解吸越有利。

4) 通过本试验可知,有机胺烟气吸收的最佳工艺条件如下:自制有机胺浓度为  $1.0 \text{ mol/L}$ , pH 值为 8, 吸收温度为  $50 \text{ }^\circ\text{C}$ , 解吸过程中最佳解吸温度为  $110 \text{ }^\circ\text{C}$ , 最佳解吸时间为  $60 \text{ min}$ 。

#### 参考文献:

- [1] 中华人民共和国环境保护部. 2010 年中国环境状况公报[EB/OL]. <http://house.ifeng.com/rollnews/detail,2011-06-03>.
- [2] 隋建才. 我国烟气脱硫技术现状与建议[J]. 能源技术(Energy Technology), 2008, 29(5): 277-280.
- [3] 韩永嘉. 烟气脱硫二氧化硫技术现状与发展趋势[J]. 过滤与分离(Journal of Filtration & Separation), 2009, 19(2): 23-27.
- [4] 刘征建. 烧结烟气脱硫技术的研究与发展[J]. 中国冶金(China Metallurgy), 2009, 19(2): 1-5.
- [5] MENG H, ZHANG S, LI C X, et al. Removal of heat stable from aqueous solutions of *N*-methyldiethanolamine using a specially designed three-compartment configuration eletrodialyzer[J]. Journal of Membrane Science, 2008, 322(2): 437-441.
- [6] 张彦锋. 国内外主流烟气脱硫技术现状及发展趋势[J]. 辽宁城乡环境科技(Liaoning Urban and Rural Environmental Science & Technology), 2004, 24(5): 53-56.
- [7] HAGHTALAB A, SHOJAEIAN A. Modeling of acid gases in alkanolamines using the nonelectrolyte Wilson-nonrandom factor mode[J]. Fluid Phase Equilibria, 2010, 289(1): 6-14.
- [8] TANG Z G, XU W Q, ZHOU C C, et al. A nonequilibrium stage model to simulate the chemical absorption of  $\text{SO}_2$ [J]. Industrial & Engineering Chemistry Research, 2006, 45: 704-711.
- [9] JESSICA L A, JANEILLE K D, EDWARD J M. Measurement of  $\text{SO}_2$  solubility in ionic liquids[J]. The Journal of Physical Chemistry B, 2006, 110(31): 1 559-1 562.
- [10] 翁淑容. 有机胺湿法烟气脱硫试验研究[D]. 南京: 南京理工大学, 2007.
- [11] 刘金龙. 可再生烟气脱硫吸收剂及工艺研究[J]. 炼油设计(Petroleum Refinery Engineering), 2002, 32(8): 37-40.
- [12] 王伟峰, 张亚通, 李达志. 新型可再生有机吸收剂“秦治-1 号”脱除烟气中二氧化硫工艺技术的研究[J]. 除尘·气体净化(Precipitation and Gas Cleaning), 2010(5): 16-19.
- [13] 王智友. 有机胺烟气脱硫现状[J]. 云南冶金(Yunnan Metallurgy), 2009, 38(1): 39-41.
- [14] 张 龙. 气体脱硫过程分析方法[M]. 北京: 化学工业出版社, 2006.
- [15] 周长城, 汤志刚. 乙二胺 / 磷酸溶液吸收  $\text{SO}_2$  的试验研究[J]. 精细化工(Fine Chemicals), 2003, 20(8): 142-147.
- [16] 杨会龙, 刘宝友, 王园园. 氨基功能化离子液体表征及吸收  $\text{SO}_2$  的实验研究[J]. 河北科技大学学报(Journal of Hebei University of Science and Technology), 2011, 32(3): 220-224.

#### (上接第 153 页)

#### 参考文献:

- [1] 刘 芳. 基于 CBR 的智能决策支持系统研究与应用[D]. 兰州: 兰州大学, 2008.
- [2] 陆汝钤. 世纪之交的知识工程与知识科学[M]. 北京: 清华大学出版社, 2001.
- [3] 陈文伟. 决策支持系统及其开发[M]. 北京: 清华大学出版社, 2000.
- [4] 周 勇, 贾瑞玉. 范例推理在智能决策系统中的应用研究[J]. 电脑知识与技术(学术交流)(Computer Knowledge and Technology(Academic Exchange)), 2007(3): 824-829.
- [5] 王 伟, 许云峰, 高 凯. 基于哈希表的动态向量降维方法的研究及应用[J]. 河北科技大学学报(Journal of Hebei University of Science and Technology), 2011, 32(4): 360-365.

#### (上接第 160 页)

- [4] SARKAR B B, CHAKI N. High level net model for analyzing agent base distributed decision support system [A]. International Association of Computer Science and Information Technology - Spring Conference (IACSITSC) [C]. Singapore: [s. n.], 2009. 351-358.
- [5] ZAKARIA N, COGBURN D L. A culturally-attuned distributed decision making model of global virtual teams in world summit on the information society [A]. The 44th Hawaii International Conference on System Sciences (HICSS) [C]. Hawaii: [s. n.], 2011. 1-10.
- [6] SUN X L, HUANG M, WANG X W. The distributed decision making risk management model for virtual enterprise based on principal-agent theory [A]. Chinese Control and Decision Conference (CCDC) [C]. Mianyang: [s. n.], 2011. 467-472.
- [7] CHENG L, HOU Z G, TAN M, et al. Necessary and sufficient conditions for consensus of double-integrator multi-agent systems with measurement noises [J]. IEEE Transactions on Automatic Control, 2011, 56(8): 1 958-1 963.
- [8] XU H K, GUO J, ZENG H T, et al. Study on distributed cooperative maintenance decision supporting system for hydropower plant [A]. IEEE International Conference on Systems, Man and Cybernetics [C]. Montréal: [s. n.], 2007. 2 296-2 301.