面向 TRIZ 的专利自动分类研究*

胡正银[1,2] 方 曙[1] 文 奕[1] 张 娴[1,2] 梁 田[1]
[1](中国科学院成都文献情报中心 成都 610041)
[2](中国科学院大学 北京 100049)

摘要: 【目的】通过构建个性化分类体系, 研究面向 TRIZ 应用的专利自动分类方法。【方法】基于主题模型, 从宏观、中观、微观三个层面构建面向 TRIZ 个性化分类体系; 通过对不同分类特征项与算法进行组合, 挑选分类准确率最高的组合构建初始分类器; 采用平滑非平衡数据与特征项降维方式对分类器进行优化, 完成对专利的自动分类。【结果】实现半自动构建面向 TRIZ 的个性化分类体系及基于该分类体系的专利自动分类。在中等数据量级场景下(千条), 实现专利自动分类, 分类效果综合评价指标高达 90.2%。【局限】该方法不适用于数据量较小(百条)时的专利分类; 在较大数据量(万条)场景下, 该方法的有效性尚未得到验证。【结论】对中等规模专利数据, 能快速构建面向 TRIZ 的分类体系, 并实现自动分类。

关键词: 发明问题解决理论 主题模型 专利分类 个性化分类体系
分类号: G353.1

1 引 言

TRIZ 是发明问题解决理论(Theory of Inventive Problem Solving)的俄文缩写, 它是 Altshuller 等分析 200 多万份专利, 归纳出关于发明具有共性的原则与方法[1]。技术矛盾(Technical Contradictions)与创新原则(Inventive Principles)是 TRIZ 核心内容之一。TRIZ 中的技术矛盾指发明中解决的各种技术难题(Problems), Altshuller 等将其归纳成 1 201 个标准工程技术矛盾; 解决这些难题的具体技术方案(Solutions), 被抽象成 40 个标准解, 即创新原则[1]。

专利分类既是组织管理专利的一种手段, 也是专利技术挖掘的重要应用场景。面向 TRIZ 的专利分类可帮助用户快速发现采用了相似发明原理或解决了相似技术难题的专利, 促进专利有效利用[2]。但直接利用现有 TRIZ 技术矛盾与创新原则进行分类, 存在一些不足。

(1) 它们主要是依靠分析机械、工程类专利得出的结果, 无法很好反映信息技术、生命科学等领域的专利特性;

(2) 它们是从专利中总结出来的一般性原则与规律, 难以直接映射到某一具体专利;

(3) 技术矛盾与创新原则长时间结构保持稳定, 描述过于抽象, 难以描述微观层面、特定领域专利集的技术特征[3]。

为此, 本文提出一种通过半自动构建个性化分类体系, 进行面向 TRIZ 的专利分类方法。该方法基于主题模型, 从技术范畴(Tech)、技术难题与解决方案(Problems & Solutions, P&S)、SAO(Subject-Action-Object)基础语义单元三个层面构建个性化分类体系, 可实现个性化、深层次、高效率专利自动分类。

2 研究背景

2.1 专利分类

专利分类是通过机器学习少量人工分好类的专利分类规则, 然后基于该规则, 将大量专利分入相应分类体系的过程。根据分类体系不同, 现有研究可分为:

通讯作者: 胡正银, ORCID: 0000-0002-5699-9891, E-mail: huzy@clas.ac.cn。

*本文系中国科学院知识产权专项工作项目“中国科学院知识产权信息服务”(项目编号:KFJ-EW-STS-032)和中国科学院西部之光项目“基于本体的专利文献技术挖掘系统研究与实践”的研究成果之一。

面向分类号、面向 TRIZ 与面向个性化分类体系分类三种[4]。

专利分类号是专利领域非常权威并且应用广泛的分类体系。分类号从技术领域的角度，采用等级的形式对专利进行分类，如国际专利分类号(International Patent Classification, IPC)将技术内容分为：部、分部、大类、小类、大组、小组，逐级形成完整的分类体系[5]。专利分类号层次结构复杂，以专家人工分类为主。

面向 TRIZ 的分类关注专利特有的 P&S 信息，它可帮助用户发现面对不同技术难题采用了相同解决方案或同一个技术难题采用不同解决方案的专利。这些专利在技术领域上可能相差很远，分布在不同的分类号中[4]。目前，研究集中在基于 TRIZ 创新原则分类。He 等[6]利用句法信息，采用关联规则进行面向创新原则的专利分类。梁艳红等[7]将创新原则归纳为显性原则与隐性原则两类，实现了面向显性创新原则的自动分类。

分类号与 TRIZ 创新原则都存在过于宽泛与抽象等不足。面向个性化分类体系的专利分类成为研究热点。Teichert 等[8]提出了基于专利功能类别构建分类体系的方法。Hu 等[9]结合主题模型与主成分分析，研究了自动构建个性化专利知识组织体系。

总之，分类号是标注专利技术领域的重要分类体系，但过于宽泛，没有反映专利特有的 P&S 信息。面向 TRIZ 专利分类是挖掘专利特有 P&S 信息的重要方法，但现有研究集中在基于抽象创新原则分类，不能满足领域个性化专利分类需求。而面向个性化分类体系专利分类研究多基于关键词构建分类体系，没有与 TRIZ 应用结合起来。

2.2　主题模型

主题模型是一系列基于概率模型、旨在发现大规模文档中隐性主题结构方法的统称。它通过分析文档集合中术语共现的概率分布，来挖掘文档集中潜在的主题及主题的概率分布[10]。LDA(Latent Dirichlet Allocation)是一个常用的主题模型，它将文档视作一系列主题的概率分布，而主题则视作一系列术语的概率分布[11]。LDA 模型已广泛应用于专利文献分析与挖掘中，相对直接将专利文献表示成关键词或 SAO 向量，LDA 生成了新的技术特征，更能揭示专利深层次知识结构，适用于构建复杂的、个性化的专利分类体系[9]。

3　研究框架与方法

鉴于现有研究的不足，本文采用 LDA 主题模型，从技术范畴、技术难题与解决方案、SAO 基础语义单元三个层面构建面向 TRIZ 的个性化分类体系，并基于该分类体系对专利进行自动分类。三个层面概念说明如表 1 所示：

表 1　面向 TRIZ 个性化分类体系概念说明

比较项	SAO	P&S	Tech
概念层级	微观层面	中观层面	宏观层面
信息类型	显性性息	隐性信息	隐性信息
概念说明	SAO 以主–谓–宾的形式直观显示专利所包含的具体技术信息	意义相近，关系密切一组 SAO 归纳成一对通用 P&S	多个专利中频繁共现的 P&S，抽象成更具一般意义的技术范畴

该方法流程如下：构建待分类领域专利数据集；基于 SAO，采用 LDA 模型构建面向 TRIZ 的个性化分类体系；基于该分类体系对专利自动分类；优化分类结果。具体流程如图 1 所示。

3.1　构建待分类领域专利数据集

针对具体技术领域，制定专利检索策略。选择合适数据源、时间段与数据过滤策略，进行专利检索，构建专利数据集。

3.2　构建面向 **TRIZ** 个性化分类体系

分类体系是自动分类的基础与前提，构建步骤如下：

(1) SAO 自动抽取与清洗

表 2　SAO 清洗步骤

清洗内容	清洗工具	作用
概念规范化	TDA[15]+领域词表	对 SAO 中 Subject、Object 同一概念的不同表达进行规范化处理
词义消歧	TDA[15]+语义词典	利用自建领域语义词典，对 SAO 中多义词进行词义消歧
功效消歧	TDA[15]+功效词表	利用领域功效概念集，对 SAO 中 Action 进行功效消歧
数据裁剪	TDA[15]	利用 TF-IDF 模型，裁剪低频 SAO
语态合并	Stanford NLP Tools[16]	有些 SAO 表达意义完全相同，但分别用主动语态、被动语义表示，将这些 SAO 予以合并

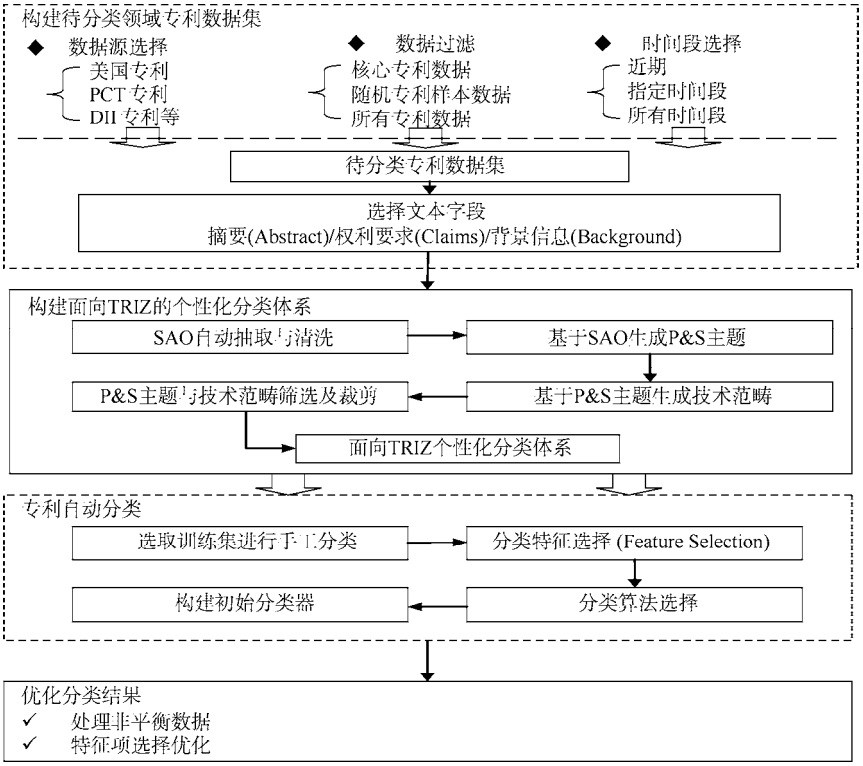

图 1　面向 TRIZ 专利自动分类流程

SAO 是一种采用主–谓–宾形式表示的三元组, 是该分类体系的基础语义单元。利用关系抽取工具如 TextRunner [12]、ReVerb [13]等从专利文本字段, 如摘要(Abstract)、权利要求(Claims)、背景知识(Background)中抽取原始 SAO。

原始 SAO 数量庞大, 表达不规范, 需要进行数据清洗。参考 Zhang 等[14]提出术语收敛(Term Clumping)框架, SAO 清洗步骤如表 2 所示。

清洗完毕后, 每一篇专利表示成一系列 SAO 组成的词袋子。

(2) 基于 SAO 生成 P&S 主题

SAO 从微观层面描述了专利包含的具体技术信息, 意义相近的一组 SAO 可归纳成一种通用的技术手段或功效, 即中观层面 P&S 主题。基于 SAO 词袋子, 采用 LDA 主题模型, 可自动挖掘出潜在的 P&S 主题。本次 LDA 的输入是直接表示 SAO 词袋子的专利——SAO 矩阵, 通过对 SAO 降维, 生成一系列 P&S 主题。经过本次主题建模, 专利表示成一系列 P&S 主题的概率分布, P&S 主题则表示成一系列语义相近 SAO 的概率分布。

定义 PS_{set} 为 P&S 主题集合, ps_j 是某一具体 P&S 主题:

$$PS_{set} = (ps_0, \cdots ps_j, \cdots ps_n) \quad (1)$$

$$ps_j = \sum_{i=1}^{1} SAO_i \cdot p(SAO_i \mid ps_j) \quad (2)$$

其中, $p(SAO_i \mid ps_j)$ 是SAO_i在ps_j中的条件概率。

有两种利用 LDA 的方式: 学习模式, 即直接从文档集中挖掘隐含的主题分布; 推理模式, 即通过学习已存在的 LDA 训练模型, 推导出新文档集的主题分布[17]。

一般来说, 推理模式适用于数量比较大的场景, 它的准确率比学习模式高。由于 SAO 词袋子数目较大, 本次 LDA 建模采用推理模式, 即先挑选出部分核心专利作为训练集, 运行 LDA 学习模式, 得到 LDA 训练模型; 然后针对所有数据, 在训练模型基础上, 运行推理模式, 得到所有专利文献关于 P&S 主题概率分布及 P&S 主题关于 SAO 的概率分布。

(3) 基于 P&S 生成技术范畴

利用 LDA 对 P&S 主题挖掘, 可自动生成一系列更宽泛主题, 这些主题可抽象成宏观层面技术范畴。本次 LDA 的输入表示成 P&S 概率分布专利——P&S

矩阵，通过对 P&S 主题降维，生成一系列技术范畴。经过本次主题建模，专利表示成一系列技术范畴的概率分布，技术范畴则表示成一系列有关联的 P&S 主题的概率分布。

定义 $Tech_{set}$ 为技术范畴集合，$tech_j$ 是某一具体技术范畴：

$$Tech_{set} = (tech_0, \cdots tech_j, \cdots tech_m) \quad (3)$$

$$tech_j = \sum_{i=1}^{k} ps_i \cdot p(ps_i \mid tech_j) \quad (4)$$

其中，$p(ps_i \mid tech_j)$ 是ps_i在$tech_j$中的条件概率。

由于 P&S 数量较少，本次 LDA 建模采用学习模式直接得到技术范畴关于 P&S 的概率分布。

(4) 主题筛选及裁剪

LDA 自动生成的 P&S 主题、技术范畴含有噪音，需要清洗之后，才能成为分类体系有效成分。主要通过两种方式清洗：将条件概率小于指定阈值的 P&S 主题与技术范畴裁剪掉；请领域专家进一步筛选有效主题。

最后，请专家给 P&S 主题和技术范畴撰写有意义的标签，得到完整的面向 TRIZ 的分类体系。与已有分类体系如专利分类号系统相比较，该分类体系有以下特点：

①该分类体系是一种个性化的分类体系。它是基于特定领域专利数据、面向具体应用的个性化分类体系；领域不同，应用目的不同，分类体系会随之动态变化；而专利分类号是面向整个技术领域、相对静态的通用分类体系[5]。

②该分类体系侧重于揭示专利的具体技术难题与解决方案信息，如它的微观层 SAO 是以“动词+名词”的形式表示某一具体的技术手段或功效；而专利分类号是按与发明创造有关知识领域进行分类[5]。

③该分类体系的描述粒度更细致、更专业；而专利分类号体系在有些领域，尤其是高技术领域，对技术分类过于宽泛，难以满足需求[5]。

该分类体系示意图如图 2 所示：

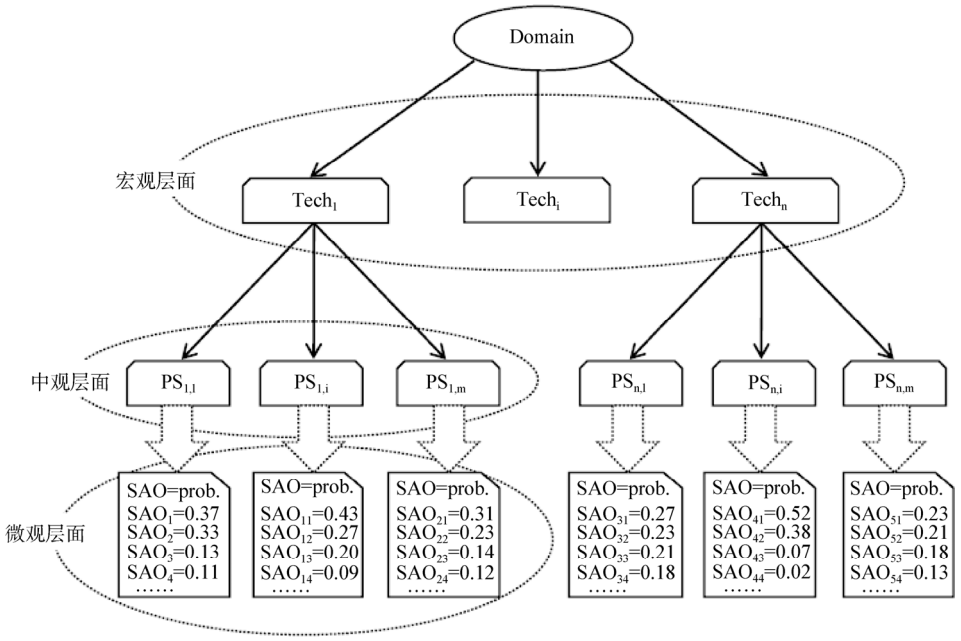

图 2　面向 TRIZ 的专利个性化分类体系

3.3　面向 TRIZ 的专利自动分类

(1) 选取训练集进行手工分类

在专利数据集中，选取一定比例的专利作为分类训练集，由专家对这些专利进行手工分类。

(2) 分类特征选择

特征选择是指从原始候选特征集中挑选或提取与任务最相关的特征集。原始候选特征集一般维度很高，相互之间存在依赖关系。如果直接利用它们进行分类，会导致分类准确率低，甚至失败。通过特征选择，将高维的原始特征集投射到低维的选定特征集中，可提高分类器的准确性与效率，是自动分类的关键步骤[18]。有两种特征选择的方式：特征提取与特征子集选取。

前者通过对原始特征集进行组合或变换，产生新的低维特征，如主成分分析、主题模型等；后者则是直接从原始特征集中挑选与任务高度相关的特征[18]。

本研究中，原始候选特征集是清洗后的 SAO 集合，采用特征子集选取方式进行特征选择。分别采用 SAO 的信息增益(Information Gain, IG)与它们在文档集中出现频率(Document Frequency, DF)作为量化 SAO 特征重要性的方法，挑选具有高特征值的 SAO 作为分类特征项。

(3) 分类算法选择

分类算法通过学习训练集，发现分类规律，进而利用该分类规律对未知数据进行自动分类。现有分类算法很多，如决策树、原生贝叶斯、人工神经网络、最大熵模型、K-近邻、支持向量机和基于关联规则的分类等[18]。每一种分类算法都有各自的特点与适用场景，需选择不同分类算法进行试分类。

(4) 构建初始分类器

通过对不同分类特征项与算法进行组合，挑选分类准确率最高的组合来构建初始分类器，利用该分类器对专利进行自动分类。

3.4 优化分类结果

为了更准确分类，需要对分类结果进行优化。本文从非平衡数据处理、SAO 特征项降维两个方面进行优化。

非平衡数据是指数据集中某几类样本数量远大于其他类，其广泛存在于各种分类问题中。如果不对非平衡数据进行处理而直接分类，分类器会将少数类样本错分到多数类，导致分类性能急剧下降。如何处理非平衡数据是数据挖掘领域的热点与难点[19]。本文通过对训练集进行多次人工重采样，以改变训练集的分布，降低不平衡性。

清洗后的SAO集合语义发散，存在很多特征相近的SAO，有利于生成P&S主题，但不利于特征选择。本文根据P&S主题关于SAO的概率分布，在特征选取前，预先对SAO进行降维，提高特征选择的准确性。

4 研究过程

选择大口径光学元件(Large Aperture Optical Elements, LAOE)专利进行面向 TRIZ 专利自动分类实证研究。

4.1 构建 LAOE 领域专利数据集

选择德温特专利(Derwent Innovations Index, DII)作为专利检索数据库，时间跨度为 2000 年-2011 申请年。由情报分析专家与领域专家共同制定检索策略，具体如表 3 所示，共得到 1 364 条专利。

表 3 LAOE 专利检索策略

序号	检索策略
# 1	TS=(Interferomet* or Interferometry or (sub-aperture stitch*)) and TS=(Interferomet* or Interferometry or SSI) and TS=(shape* or surface* or profil* or "Surface precision" or topograph) database =CDerwent, EDerwent, MDerwent Timespan =2000-2011
# 2	IP=G01B-009/02 or IP=G01B-009/021 or IP=G01B-009/023 or IP=G01B-009/025 or IP=G01B-009/027 or IP=G01B-009/029 database =CDerwent, EDerwent, MDerwent Timespan=2000-2011
# 3	#1 and #2 database =CDerwent, EDerwent, MDerwent Timespan =2000-2011

4.2 构建 LAOE 领域面向 TRIZ 的个性化分类体系

(1) SAO 自动抽取与清洗

采用开放式实体关系抽取工具 ReVerb[13]从 DII 数据中的 Title、Abstract 字段中抽取 SAO，共得到 20 957 条原始 SAO。利用表 2 所示清洗步骤进行 SAO 清洗，得到 4 892 条 SAO。由具有领域背景的分析人员进一步筛选，并省略部分 Subject/Object 单元，得到 2 372 条 SAO 基础语义单元，如表 4 所示:

表 4 基础语义单元 SAO 示例

序号	SAO
1	**check** large lens convex surface
2	**measure** surface roughness
3	method **analyze** interference-fringe
4	device **measure** lens deflection
5	**monitor** surface quality
…	……

(2) 基于 SAO 生成 P&S 主题

采用机器学习工具集 MALLET[17]中 LDA 模块对 SAO 进行降维，生成 P&S 主题。LDA 的重要参数包括“主题数目”、“Dirichlet 先验参数”等。参数值设置不合适，会导致最终分类体系过于宽泛或过于狭窄。经

反复实践，本文采用的参数设置原则为：绝大部分专利能被 10%的主题表示；且绝大部分主题能被 1%的 SAO 表示[9]。部分 LDA 重要参数配置如表 5 所示。在这组参数配置下，84.16%(1 148)的专利能被 10%(20)的主题表示；83.50%(167)的主题能被约 1%(20)的 SAO 表示。

表 5　LDA 参数配置

参数名称	参数值	参数含义
num-topics	200	P&S 主题数目
num-iterations	1000	Gibbs 采样迭代次数
alapha	0.25	估算 P&S 分布的 Dirichlet 先验参数
beta	0.01	估算 SAO 分布的 Dirichlet 先验参数

通过 LDA，得到 1364×200 的专利——P&S 主题分布矩阵及 200×2372 的 P&S 主题——SAO 分布矩阵。矩阵的权重为相应的条件概率值。

(3) 基于 P&S 生成技术范畴

进一步采用 LDA 学习模式对 P&S 主题进行降维，生成 Tech 主题。LDA 参数配置中，Tech 主题数目设置为 20，其他参数保持不变。得到 1364×20 的专利——Tech 主题分布矩阵及 20×200 的 Tech 主题——P&S 主题分布矩阵。矩阵的权重为相应的条件概率值。

(4) 主题筛选及裁剪

裁剪掉权重较低的主题。设置 Tech 主题的阈值为 0.1，即剔除专利——Tech 主题分布矩阵中权重小于 0.1 的 Tech 主题；设置 P&S 主题的阈值为 0.05，即剔除 Tech 主题——P&S 主题分布矩阵中权重小于 0.05 的 P&S 主题。由情报分析专家与领域专家共同对主题进一步筛选、合并。得到 124 个有效 P&S 主题与 4 个有效 Tech 主题。LAOE 领域的面向 TRIZ 的个性化分类体系如表 6 所示。

4.3　面向 TRIZ 的专利自动分类

(1) 选取训练集进行手工分类

选取 100 条专利作为训练集。请专家人工将这 100 条专利归入{C1,C2,C3,C4}类中。为了尽可能保证数据平衡，按分类号分布比例挑选训练集。

(2) 分类特征选择

分别选择 top5、top10、top20 IG 与 DF 大于阈值 2、3、5 的 SAO 作为分类的特征项。

(3) 分类算法选择

选择最大熵模型(Maximum Entropy Classifier, MaxEnt)、决策树分类(C4.5 Decision Tree Classifier, DT)、原生贝叶斯分类(Naïve Bayes, NB)三种分类算法，利用 MALLET[17]工具集中分类模块进行试分类。

表 6　LAOE 领域的面向 TRIZ 的个性化分类体系

序号	技术范畴	P&S 主题	SAO
C1	Measuring Surface Shape	$P\&S_{10}$(p=0.557) $P\&S_{11}$(p=0.213) $P\&S_{14}$(p=0.117) ……	**check** large lens convex surface; **measure** surface roughness; **analyze** object surface profile; ……
C2	Surface Measuring Method	$P\&S_{1}$(p=0.628) $P\&S_{39}$(p=0.124) $P\&S_{114}$(p=0.017) ……	method **measure** diffraction; method **measure** optical curvature; method **analyze** interference-fringe; ……
C3	Surface Measuring Device	$P\&S_{15}$(p=0.415) $P\&S_{79}$(p=0.354) $P\&S_{102}$(p=0.203) ……	device **measure** wave aberration; **shear** interferometer for flatness; device **measure** lens deflection; ……
C4	Online Monitoring	$P\&S_{27}$(p=0.813) $P\&S_{42}$(p=0.102) $P\&S_{78}$(p=0.005) ……	**monitor** surface quality; **control** optical surface quality; **inspect** surface shape; ……

(4) 构建初始分类器

选取 300 条专利，平均分成三组$\{t_1, t_2, t_3\}$。交叉选取$\{t_1, t_2\}$、$\{t_1, t_3\}$、$\{t_2, t_3\}$作为测试集。在不同分类特征项下，三种分类算法在三组测试集的平均分类准确率如表 7 所示：

表 7　三种分类算法在不同特征项下分类准确率

特征项	MaxEnt 准确率(%)	DT 准确率(%)	NB 准确率(%)
IG (top5)	67.6%	74.6%	72.6%
IG (top10)	73.5%	**82.8%**	80.5%
IG (top20)	71.3%	77.2%	79.1%
DF(threshold=2)	57.2%	65.2%	71.2%
DF(threshold=3)	52.6%	68.9%	69.3%
DF(threshold=5)	59.3%	72.7%	74.8%

分类器准确率越高，表示对待分类专利数据自动分类的能力越强。从表 7 可以看出，当选择 top 10 IG 的 SAO 作为分类特征项、DT 作为分类算法时，分类

准确率最高，达 82.8%。因此，选择该组合构建初始分类器，对全部 LAOE 专利分类。

4.4 优化分类结果

常用评价分类效果指标包括：准确率 P (Precision)、召回率 R (Recall) 与综合评价指标 F-measure。P 表示已分类数据中分类正确的比例[20]；R 表示已正确分类数据占所有应该被分到该类数据的比例[20]；F-measure 是综合考虑 P 和 R 性能的指标，常用的 F-measure 为 F1，F1 = 2PR/(P+R) [20]。

基于初始分类器分类效果欠佳，需要优化。本文从非平衡数据处理、SAO 特征项降维两方面优化。首先人工重采样时，通过增加少数类样本数量(Over-Sampling)与减少多数类样本数量(Under-Sampling)来处理非平衡数据。然后特征选择时，通过裁剪低概率 SAO 等方式对其降维，提高特征选择准确性。最后技术范畴分类效果如表 8 所示：

表 8　专利技术范畴分类准确率

序号	P	R	F1
C1	0.884	0.82	0.851
C2	**0.926**	**0.88**	**0.902**
C3	0.782	0.78	0.781
C4	0.726	0.86	0.787

5 结果与讨论

本文基于 LDA 主题模型，提出一种面向 TRIZ 的专利自动分类方法。基于 LAOE 专利实证研究发现：LAOE 专利集中在“光学元件面形检查(Measuring Surface Shape)、面形测量方法(Surface Measuring Method)、面形测量装置(Surface Measuring Device)、在线面形监测(Online Monitoring)” 4 个技术范畴。这 4 个技术范畴包涵“检查大透镜的凸面(Checking Large Lens Convex Surface)、衍射法测量大光学元件曲率(Diffraction Method for Measuring Large Optical Curvature)、波像差测量装置(Wave Aberration Measuring Device)、光学元件表面质量控制(Optical Surface Quality Control)”等 124 个具体技术问题。这些技术问题又可用 2 372 个 SAO 来进行描述。基于该分类体系对 LAOE 专利进行分类，技术范畴的分类准确率最高达 92.6%，最低为 72.6%；召回率最高达 88%，最低为 78%；综合分类评价参数 F1 值最高达 90.2%，最低为 78.1%。目前，基于该方法构建的 LAOE 专利辅助创新知识库系统已成功在中国科学院上海光学精密机械研究所部署应用。

该方法在实际应用中还存在一些问题：

(1) 将中观层面的 P&S 主题看成一个整体考虑，没有进一步区分成 Problems 与 Solutions 主题及自动发现它们之间的语义关系。如何将 SAO 定向、准确生成 Problems 与 Solutions 主题，并挖掘出它们之间的语义关系，是未来研究的难点与重点。

(2) 在数据量较小(如几百条)时，该方法效果较差，与 LDA 对数据量有一定要求有关；在较大数据量级(上万条)场景下，该方法的有效性尚未得到验证。

(3) 虽然领域专家参与是专利技术挖掘必不可少的部分，但如何规范领域专家参与模式、减少其主观影响，增强该方法的稳定性与通用性，也是未来研究重点。

6 结　语

本文基于LDA主题模型与SAO基础语义单元，提出了一种面向TRIZ的专利自动分类方法。该方法实现了半自动构建面向TRIZ个性化分类体系及基于该分类体系的专利自动分类。实证研究表明：在中等数据量级场景下(千条)，该方法可高效、准确地实现面向TRIZ专利自动分类。基于该方法构建的面向TRIZ个性化分类体系还可应用于更多的专利技术挖掘领域，如专利语义检索、技术演化分析、发现核心专利、热门专利预测等。

未来，笔者将在较大数据量级场景下(万条)，对该方法的有效性进行研究；自动区分与挖掘 Problems 与 Solutions 之间的关系，也是未来研究的重点。

参考文献：

[1] Kaplan S. An Introduction to TRIZ: The Russian Theory of Inventive Problem Solving [EB/OL]. [2013-07-02]. http://www.trizasia.com/FileStorage/634166595685730035 2005-Intro_to_TRIZ%20--%20for%20printer.pdf.

[2] Loh H T , He C, Shen L. Automatic Classification of Patent Documents for TRIZ Users [J]. World Patent Information, 2006, 28(1): 6-13.

[3] Hu Z Y, Fang S, Liang T. Automatic Patent Classification Oriented to Problems & Solutions [C]. In: Proceedings of

Conference on Artificial Intelligence and Data Mining (AIDM'13), Sanya, China. 2013: 22-24.

[4] 胡正银, 方曙. 专利文本技术挖掘研究进展综述[J]. 现代图书情报技术, 2014(6): 62-70. (Hu Zhengyin, Fang Shu. Review on Text-based Patent Technology Mining [J]. New Technology of Library and Information Service, 2014(6): 62-70.)

[5] WIPO. International Patent Classification (Version 2014) [EB/OL]. [2014-06-01]. http://www.wipo.int/export/sites/www/classifications/ipc/en/guide/guide_ipc.pdf.

[6] He C, Loh H T. Pattern-oriented Associative Rule-based Patent Classification [J]. Expert Systems with Applications, 2010, 37(3): 2395-2404.

[7] 梁艳红, 檀润华, 马建红. 面向产品创新设计的专利文本分类研究[J]. 计算机集成制造系统, 2013, 19(2): 382-390. (Liang Yanhong, Tan Runhua, Ma Jianhong. Study on Patent Text Classification for Product Innovative Design [J]. Computer Integrated Manufacturing Systems, 2013, 19(2): 382-390.)

[8] Teichert T, Mittermayer M A. Text Mining for Technology Monitoring [C]. In: Proceedings of 2002 IEEE International Engineering Management (IEMC'02). IEEE, 2002: 596-601.

[9] Hu Z, Fang S, Liang T. Empirical Study of Constructing a Knowledge Organization System of Patent Documents Using Topic Modeling [J]. Scientometrics, 2014, 100 (3): 787-799.

[10] Blei D M. Probabilistic Topic Models [EB/OL]. [2013-06-12]. https://www.cs.princeton.edu/~blei/kdd-tutorial.pdf.

[11] Blei D M, Ng A Y, Jordan M I. Latent Dirichlet Allocation [J]. Journal of Machine Learning Research, 2003, 3: 993-1022.

[12] Yates A, Cafarella M, Banko M, et al. TextRunner: Open Information Extraction on the Web [C]. In: Proceedings of NAACL-Demonstrations '07 of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations. Association for Computational Linguistics, 2007: 25-26.

[13] Fader A, Soderland S, Etzioni O. Identifying Relations for Open Information Extraction [EB/OL]. [2013-03-02]. http://ai.cs.washington.edu/www/media/papers/reverb.pdf.

[14] Zhang Y, Porter A L, Hu Z, et al. "Term Clumping" for Technical Intelligence: A Case Study on Dye-sensitized Solar Cells [J]. Technological Forecasting and Social Change, 2014, 85: 26-39.

[15] Thomson Reuters. Thomson Data Analyzer [EB/OL]. [2013-03-03]. http://ip-science.thomsonreuters.com.cn/media/tda.pdf.

[16] The Stanford Natural Language Processing Group. Research [EB/OL]. [2013-03-03]. http://www-nlp.stanford.edu/research.shtml.

[17] Mimno D. Machine Learning with MALLET [EB/OL]. [2013-03- 03]. http://mallet.cs.umass.edu/mallet-tutorial.pdf.

[18] 杨建武. 文本自动分类技术 [EB/OL]. [2013-06-13]. http://www.icst.pku.edu.cn/course/mining/11-12spring/TextMining04-%E5%88%86%E7%B1%BB.pdf. (Yang Jianwu. Review on Text Classification [EB/OL]. [2013-06-13]. http://www.icst.pku.edu.cn/course/mining/11-12spring/TextMining04-%E5%88%86%E7%B1%BB.pdf.)

[19] 钱洪波, 贺广南. 非平衡类数据分类概述[J]. 计算机工程与科学, 2010, 32(5): 85-88. (Qian Hongbo, He Guangnan. A Survey of Class-imbalanced Data Classification [J]. Computer Engineering & Science, 2010, 32(5): 85-88.)

[20] Powers D M W. Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation [EB/OL]. [2013-03-03]. http://www.infoeng.flinders.edu.au/research/techreps/SIE07001.pdf.

作者贡献声明：

胡正银: 文献调研, 实证分析, 论文撰写;
方曙: 研究命题的提出、设计, 论文修订;
文奕: LDA 主题模型应用;
张娴: 领域词表建设, 面向 TRIZ 分类体系构建;
梁田: SAO 数据清洗, 分类数据处理。

收稿日期: 2014-07-23
收修改稿日期: 2014-08-22

Study on Automatic Classification of Patents Oriented to TRIZ

Hu Zhengyin[1,2] Fang Shu[1] Wen Yi[1] Zhang Xian[1,2] Liang Tian[1]
[1] (Chengdu Document and Information Center, Chinese Academy of Sciences, Chengdu 610041, China)
[2] (University of Chinese Academy of Sciences, Beijing 100049, China)

Abstract: [**Objective**] This paper proposes an approach to automatically classify patents oriented to TRIZ applications based on a personalized classification system. [**Methods**] A personalized classification system is constructed in micro-macro-meso levels using topic model. Then, an appropriate feature and classifier are chosen to preliminarily classify patents. The classifier is optimized by smoothing unbalance data and reducing features dimensions. [**Results**] This approach implements semi-automatically constructing a personalized classification and automatically classifying patents oriented to TRIZ applications. In medium data size, this approach can classify patents with F-measure value of 90.2%. [**Limitations**] This approach is not available in small size data set and not verified in big size data set. [**Conclusions**] This paper can classify patents oriented to TRIZ applications in medium data size.

Keywords: TRIZ Topic model Patent classification Personalized classification system

《现代图书情报技术》杂志启用作者 ORCID 号

在学术论文投稿、基金申请提交、科研产出管理和全球化学术交流活动中，科研人员身份识别已成为一个关键环节。作为使用最为广泛的“科研人员身份证”——ORCID，是一套免费的、全球唯一的 16 位身份识别符。目前，世界上已有 120 余家极具影响力的出版社、基金组织以及科研机构采用 ORCID 号标识作者身份。

《现代图书情报技术》杂志于 2015 年第 1 期开始启用 ORCID 服务，对出版论文标记通讯作者的 ORCID 号。拥有 ORCID 号后，将有效助推作者的学术工作:

(1) 拥有专属的国际唯一学术识别符，获得科研身份证，作者与全球同行之间“知你、知我”;

(2) 满足投稿期刊、基金组织对作者的唯一识别要求;

(3) 即时发现分布在各处的个人科研产出。

《现代图书情报技术》杂志投稿系统已完成升级改造，作者在投稿过程中即可方便地实现 ORCID 号自动注册。此外，还可以通过 ORCID 官方网站(http://orcid.org/)或 ORCID 中国合作服务平台 iAuthor(http://iauthor.las.ac.cn)实现 ORCID 注册申请。

(本刊讯)