面向循证医学的科技文献摘要结构化表示研究

杜圣梅[1] 朱礼军[1] 徐 硕[2]
（1. 中国科学技术信息研究所，北京 100038；2. 北京工业大学，北京 100124）

摘要：临床科学研究往往以科技文献的形式储存。文章对医学领域科技文献表示模型进行概述和分析，以PIBOSO模型为基础，采用支持向量机对科技文献的摘要句子进行分类，实现了科技文献摘要信息的自动化抽取及关键句子的识别，从而将科技文献的摘要内容进行语义关系的量化和结构化表示，为临床医师和相关研究人员在寻找证据资源时提供有效借鉴和帮助。实验结果表明，该方法的F值在大多数类别上高于其他方法，表明研究方案具有可行性和有效性。

关键词：循证医学；SVM；句子分类；知识挖掘；机器学习
中图分类号：G202; TP391 **文献标识码：**A **DOI：**10.3772/j.issn.1674−1544.2018.06.013

Research on Structured Presentation of Scientific Literature Abstracts for Evidence-based Medicine

DU Shengmei[1], ZHU Lijun[1], XU Shuo[2]
(1. Institute of Scientific and Technical Information of China, Beijing 100038; 2. Beijing University of Technology, Beijing 100124)

Abstracts: It is well known that clinical scientific research is often stored in scientific and technical (S&T) literature. The knowledge hidden in S&T literature can provide clinicians and researchers with the clinical decision-making evidence in the practice of evidence-based medicine. After the representation models for S&T literature are summarized and analyzed in the medical field, are fined PIBOSO model is used in this study. For purpose of the automatic extraction of the abstract and the identification of key sentences, Support Vector Machine (SVM) is utilized here to classify abstract sentences. The classification results with SVM help quantify the semantic relations and structure the abstracts, thus providing effective reference for clinicians and researchers to find evidence. From experimental results, one can see that F-score in this work is higher than the counterparts in most categories, which indicates that our research framework is feasible and effective in the sentence classification task from the biomedical field.

Keywords: evidence based medicine, SVM, sentence classification, knowledge mining, machine learning

作者简介：杜圣梅（1993—），女，中国科学技术信息研究所硕士研究生，研究方向：信息技术与知识技术、知识工程；**朱礼军**（1974—），男，中国科学技术信息研究所研究员，研究方向：智能信息处理、知识组织与知识工程、移动问答、语义网等；**徐硕**（1979—），男，北京工业大学校聘教授，研究方向：商务智能与科技决策，大数据和数据挖掘等（通讯作者）。

基金项目：北京市社会科学基金项目“大数据驱动的可制造性知识挖掘与管理方法研究”（17GLB074）；北京市优秀人才培养资助青年骨干个人项目”（2015000020124G052）。
收稿时间：2018 年 6 月 6 日。

0 引言

循证医学（Evidence Based Medicine，EBM）[1-2]源于临床实践，对医疗模式的转变产生了巨大影响，传统以疾病为中心的生物医疗模式已经转变为以患者为中心的“现代生物—心理—社会—医疗”模式。作为一种新的医疗模式，循证医学概念的内涵和外延历经了长期的发展，已经日趋完善。其核心思想是：当临床医师在进行医疗决策时，必须以客观真实的临床科学研究为依据，并结合自身的临床专业知识和患者本人的意愿[3]。临床医师若要基于EBM做出高效判断，全面、可靠、相关和及时地获取证据至关重要。

随着全球科技的进步，知识更新的速度越来越快，生物医学领域的科技文献呈爆炸式增长。尽管这些科技文献大多都经过了同行评议，但是质量却良莠不齐。作为临床医师决策时至关重要的证据通常埋藏于海量的科技文献中。因此，新的医疗模式对临床医师及研究人员提出了新的要求：（1）需要培养拥有持续从科学研究中学习汲取相关新知识的能力；（2）能够迅速定位相关医学文献，并能客观、准确地评价其质量和适用性，寻找到目前最佳的临床证据并最终应用到解决临床问题中。

根据循证医学的临床指南，生物医学领域科技文献的组织结构通常遵循PICO模型[4]，即Population（P）、Intervention（I）、Comparison（C）、Outcome（O）。临床医师在判定临床研究（例如随机对照试验，RCT）是否与待解决的问题相关时，也通常参考该模型或其变种。随着人工智能和机器学习的飞速发展，其方法和思想已成功应用于多个领域，并取得了良好的效果。生物医学领域科技文献所表现的PICO组织结构模式，为基于机器学习方法自动提取相关信息提供了方便。

笔者通过实际访谈发现，临床医师往往通过科技文献摘要部分的阅读，即可初步判断证据资源的相关性和有效性。因此本文尝试识别生物医学领域科技文献摘要部分的关键句子，将其映射到特定组织结构模型中的各个部分，从而实现摘要内容的建模及语义关系的量化表示，为具体实践循证医学的相关研究提供支撑。

本文其余部分的组织结构如下：第1节概述分析循证医学领域里科技文献表示模型和科技文献结构化表示的相关研究，并分析对比本文方法与其他研究的不同和进步；第2节是对科技文献建模表示所展开的具体研究，其中包括科技文献表示模型PIBOSO的概述、分类特征向量的构建和分类模型的选择与设计；第3节是实验及结果分析；第4节是总结全文。

1 相关工作

1.1 科技文献表示模型

临床医生往往需要在大量已发表的文献中定位、总结出有效信息，以便于全面客观地了解临床问题的相关状况，从而迅速明智地做出医疗决策。早期的研究集中于如何构建有效的信息检索模型，以便于临床医生及科研人员进行更有效的检索，取得更多的有效的信息。这些模型也为科技文献的自动分析和智能鉴别奠定了基础，表1中列举了现今常用于检索、归类、识别科技文献的表示模型。最初设计PICO模型主要是为了

表1 循证医学科技文献表示模型

模型	描述
PICO	Patient−Population−Problem/Intervention /Comparison /Outcome
PICOT	Patient−Population−Problem/Intervention /Comparison /Outcome /Time
PESICO	Patient−Population−Problem/Environment/Stakeholders /Intervention /Comparison /Outcome
PECODR	Patient−Population−Problem/Exposure−Intervention/Comparison/Outcome/Duration/Results
PIBOSO	Population/Intervention/Background/Outcome/Study Design/Other

辅助临床人员将临床问题转化为可进行检索的关键词，并且自从PICO被提出后，一些基于PICO的扩展模型相继被提出，如加入“时间”的PICOT模型[5]，加入“环境”和“利益相关者”的PESICO模型[6]，加入“持续时间”和“结果”的PECODR模型[7]以及加入“研究设计”和“背景”的PIBOSO模型[8]。

其中的PIBOSO作为PICO检索标准的一种扩展，其设计的初衷是为了实现已发表文献摘要的自动抽取，为应用机器学习的方法实现摘要内容的结构化表示提供了有效借鉴，因此本文决定采用该模型展开对科技文献摘要的结构化表示的研究。

1.2　科技文献摘要结构化表示方法

随着医学领域知识更新速度越来越快，生物医学相关的科技文献呈爆炸式增长，使得临床医生在海量科技文献中能够迅速找到与某临床问题密切相关的研究变得越来越困难，因此寻求一种自动识别生物医学文献句子类型的方法已经引起了众多研究人员的关注。

Demner−Fushman等[9]首次提出了基于PICO标准的句子自动分类模型，该模型通过基于规则的方法建立了句子分类器，而对Outcome类别构建特征向量，包括n元语法、位置、语义等特征，进行有监督分类。该分类器对275篇手动标注的文献摘要进行训练，其准确率达到了0.74~0.93，但局限性在于训练数据集较少且其中一些类别要依靠手动构建规则的方法。Kim等[8]基于细化的PIBOSO模型将句子分类过程分为两个阶段：第一步构建分类器识别包括PIBOSO概念的句子；第二步构建另外一个分类器将PIBOSO的标签类型分配给句子。该研究采用条件随机场（Conditional Random Field，CRF）作为分类模型，并结合医学摘要特点构建了特征向量，包括领域知识、语义、结构及顺序等特征。Sarker等[10]将多类分类问题转化为多个二分类问题，采用SVM作为分类器，构建了包括二元语法、句子位置、句子长度、节标题、领域知识和语义信息的特征向量，最终使F值达到了0.80。与上述研究不同的是，本文在构建特征向量时尝试不借用外部资源[11]，并在此基础上进行扩展，融入了句子的统计信息特征，从而进一步利用句子的模式信息对句子进行分类，以达到对摘要文本进行结构化表示的目的。

2　科技文献摘要结构化表示方法研究

2.1　科技文献摘要表示模型

科克伦协作网（Cochrane Collaboration）在制定系统评价时因其方法的科学性及严谨性，使Cochrane临床指南被誉为循证医学领域中最为权威的证据来源之一。其纳入标准综合了所提出临床问题的各个方面以及回答这些临床问题的研究类型。其中，关于防治性研究系统评价中，高质量科技文献及临床试验研究的选择纳入标准主要包含了4个方面[12]：研究设计类型、研究对象、干预措施和对照措施、结局指标。该纳入标准蕴含了对医学文献进行筛选和鉴别时的基本流程和标准。而科技文献摘要表示模型PIBOSO所包含的6种信息类型恰能将纳入标准的4个方面完全覆盖。B（Background）：交代本次研究的来源及现状。P（Population）：构成研究样本的个体、对象或者项。I（Intervention）：研究过程中改变条件或者改变流程的干预行为。O（Outcome）：总结干预措施的影响和结果。S（Study Design）:摘要中用于描述研究的部分。O（Other）：不属于上面任何一个类别，并且已经假定对临床决策提供的帮助很少，即非关键性和不相关的句子。

除此之外，医学领域文献的摘要提取的有效信息往往包含：背景、研究对象及临床特点、干预措施、结果、研究设计等。考虑到后续研究中信息抽取的可行性，本次研究选用PIBOSO模型，研究摘要句子自动分类方案，用于医学领域文献摘要的建模表示及语义关系的量化。

2.2　分类特征分析

该部分内容是基于PIBOSO文献摘要建模过程中进行句子分类时所展开的特征选择研究。本文基于对生物医学科技文献摘要的分析，确定使用以下3组特征对分类器进行训练。

（1）词汇特征

词汇特征作为句子表述的基本信息被纳入到本文的方法中。该特征主要包含词性标注和Lemma特征。词性（POS）标注特征是指每个句子中所标记的POS标签作为一个特征。标签包括名词、形容词、副词、动词、连词、介词和代词，动词的否定形式也被纳入其中。为了适应词汇的各种变形，本文将词形进行还原后作为特征纳入到方法中，并且实验数据集在预处理时已经将数据的词形进行了还原。

（2）统计特征

统计特征指的是动词和非动词的分布。Waard等[13]指出动词信息可以作为科技文献分类中一个很好的指标，此外句子的一些模式信息也在动词中得以体现，比如动词的过去时往往存在于Population和Intervention的句子类别中，而动词的否定式常出现在Outcome句子类别中。所以本文将动词的不同语态、时态以及否定形式的统计数量进行归一化后作为特征纳入到方法中。

（3）位置特征

生物医学领域科技文献的摘要叙述遵循一定的标准形式，即：阐述背景信息，陈述问题，如何解决该问题，最后以结果描述结束。该结构信息可以很好地与PIBOSO模型分类的目标类别关联。因此，摘要中句子的位置信息可以作为句子分类的特征。

本文基于上述分析，通过举例具体说明了以上特征。下面选取了PMID=10819426该文章的摘要前4句内容及每句对应的分类标签，并详细展示了第三句抽取的词汇特征、统计特征和位置特征，如图1所示。

综上所述，本文结合医学领域科技文献句子的特征，给出了用于构建分类器模型的整体特征向量，如表2所示。

2.3 基于SVM的句子分类模型

支持向量机（SVM）[14]自1995年被提出以来得到了迅速的发展。由于SVM遵循结构化风险最小的原则，使得推广泛化能力明显优于传统

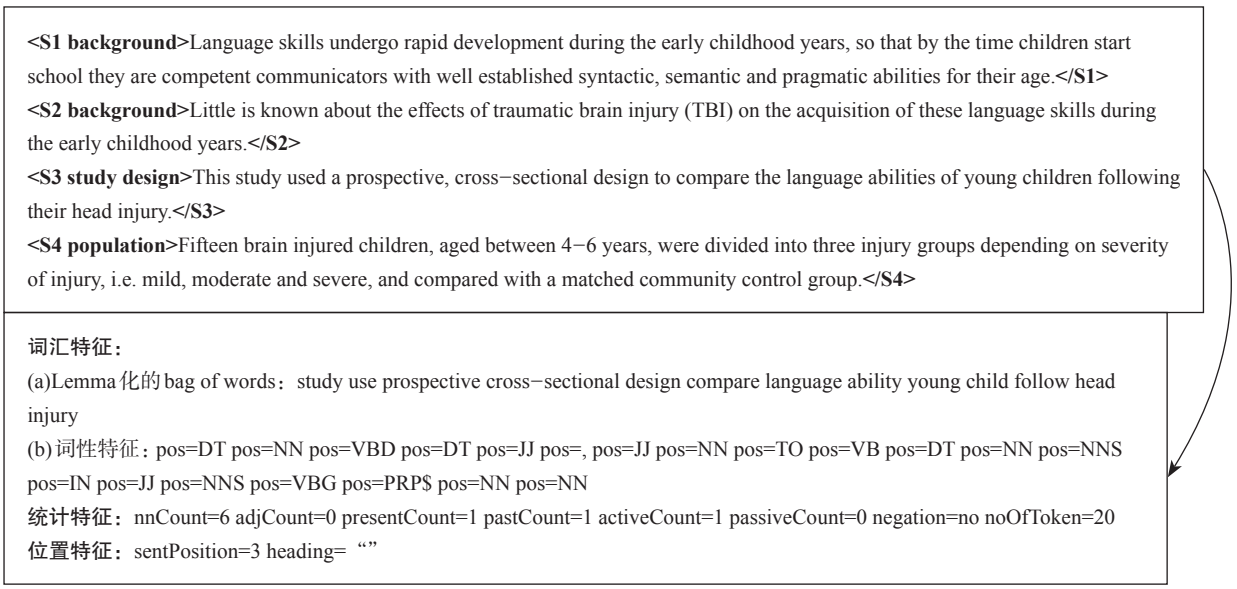

图1 摘要文本特征抽取示例

表2 训练分类模型所构建的特征向量

特征集	特征向量
词汇特征	Lemma化的词袋－词性
统计特征	名词数量－形容词数量－不同时态动词的数量－不同语态动词的数量－否定形式动词的数量－词条数量
位置特征	句子位置

机器学习算法，其核心思想是通过核函数将特征映射到高维特征空间，在此空间寻找最优的划分超平面，巧妙地将求解过程转化为著名的凸二次规划问题，保证存在全局最优解。常用的核函数通常包含四类：线性核函数、多项式核函数、径向基（RBF）核函数和Sigmoid核函数。本文选用的是RBF核函数，原因主要为以下3点：一是Sigmoid核函数是非正定的，且在对某些参数设置时，Sigmoid核函数的性能与RBF核函数差不多。二是多项式核函数由于其具有较多的参数，使得在模型选择时更为困难一些。除此之外，多项式核函数还存在数值问题，比如数值的上溢和下溢[15]。三是相比较而言，RBF核函数具有良好的光滑性，通常在缺少先验知识的情况下成为较理想的选择[16]。图2给出了面向循证医学的科技文献摘要句子分类的流程图。

3　实验及结果分析

3.1　实验数据集

本文的实验数据集为澳大利西亚语言技术协会（Australasian Language Technology Association，ALTA）2012年公布的比赛数据NICTA-PIBOSO[8, 17]，该数据集包含1000篇不同主题领域的文献摘要，由一位医学专业的学生历经80小时按照特定分类标准（PIBOSO）对摘要句子进行了标注，最终通过了一致性检验。为了训练和评测分类器，该数据集被分成两份，其中800篇已标注的文献摘要为训练集，另外200篇为测试集。图3是对数据集中各类别的句子样本数量分布的统计。

3.2　实验结果及讨论

在实验过程中，本文对上述数据集进行预处

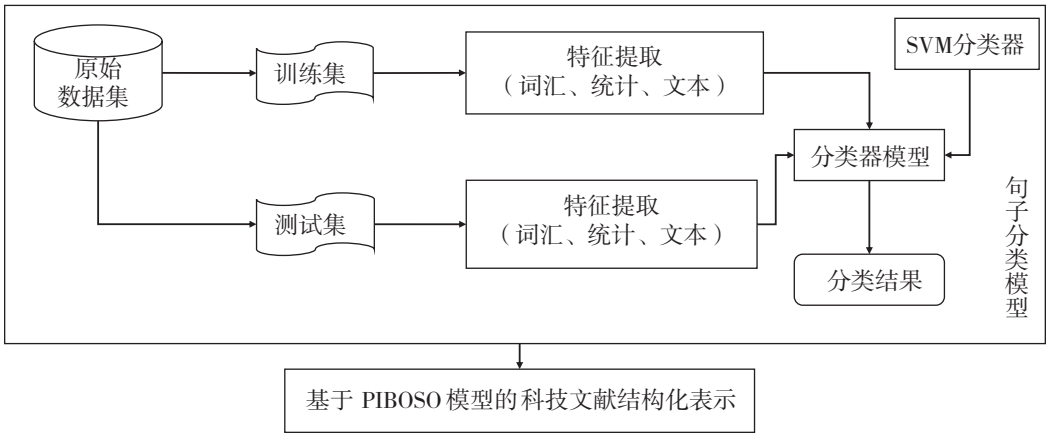

图2　面向EBM的科技文献摘要句子分类流程图

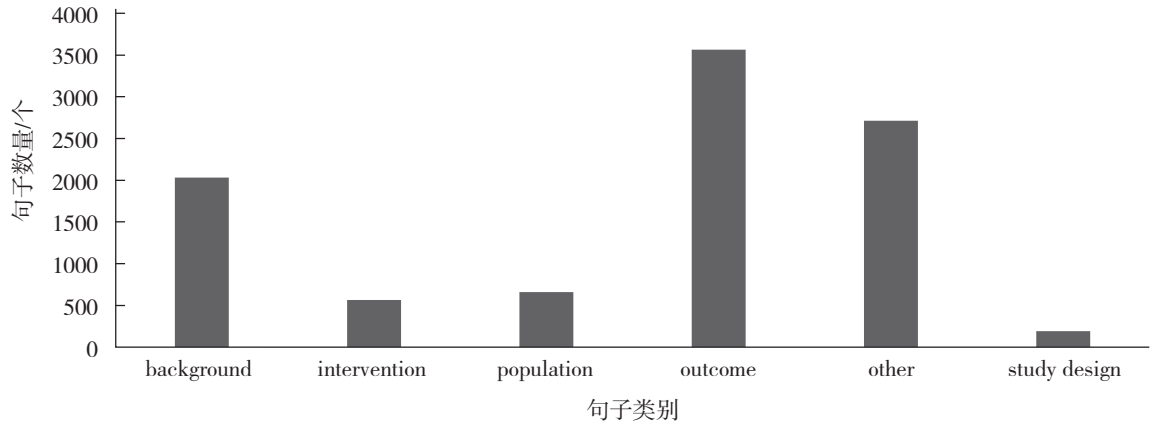

图3　NICTA-PIBOSO的句子类别样本量分布

理并构建特征向量，所应用的机器学习分类工具是LibSVM。由于本文基于PIBOSO模型对摘要句子进行分类，实质上是一个多类别分类实验，所以采用了一对多的分类方式，其中SVM分类器选用径向基函数，惩罚因子采用默认值0.001。在进行参数寻优后，结合样本数量分布，训练得到6个二分类器，并应用P、R、F作为评价分类器的指标对预测结果进行了分析。其中，P为准确率，R为召回率，F值作为综合评价指标，计算公式为$F=2\times P\times R/(P+R)$，从而可以使每个类别的分类结果更加明确直观。表3中给出了本文方法在6个类别上分类结果的准确率、召回率及F值。

由于数据集中样本数量分布差异较为明显，所以分类器在样本量较大的类别上（即Background，Outcome，Other类别）能够得到较为理想的分类结果，F值达到了0.72~0.87，但是在样本数量较少的类别上分类效果较差。本文在实验过程中通过参数寻优以及设置正负类的惩罚权重的方法对分类器进行了优化，有效地提高了分类器的准确率、召回率及F值。

除此之外，为了了解本次实验设计的分类器的性能，表4中也给出了与以往方法[8, 10]分类结果的比较，其中由于Kim等[8]在研究中将该数据集分成结构化和非结构化两部分，所以本文在比较F值时，对该研究在两部分数据集上所得的F值进行加权求均值的计算，最终得出表4中的综合F值。

我们可以看出，本文方法在大多数句子类别的分类效果上表现优于A−MQ和Kim等提出的研究方法，即Background、Intervention、Outcome和Population 4个类别，只有Other和Study Design两个类别的分类效果不太理想，但识别摘要中非关键性句子（Other类别）的F值也达到了0.72。本文方法在上述4个类别上较好的表现说明加入的统计特征在句子分类时是有效的，这在一定程度上提高了句子类别的辨识度。也从侧面说明了其他两种方法使用外部标注工具得到的医学领域知识特征，对区分句子类别，尤其是对上述4个类别的贡献较小，但在Intervention类别上表现较为显著。

4 结语

为实现科技文献摘要文本的建模表示及语义关系的量化，本文使用SVM分类器对科技文献进行知识挖掘并对摘要文本进行句子分类，构建了包含词汇、统计及位置的特征向量来训练分类模型。通过实验得出分类结果，对比现有的句子分类方法，本文方法在大多数类别上获得了较高的F值，表明了该方法的有效性。如果能进一步考虑上下文信息和数据分布的不平衡性，将会得到更好的分类结果。

表3 本文方法在各类别得到的P、R及F值

	Background	Intervention	Outcome	Population	Other	Study Design
P	0.79	0.44	0.80	0.58	0.70	0.75
R	0.81	0.49	0.94	0.57	0.74	0.35
F	0.80	0.46	0.87	0.57	0.72	0.47

表4 本文方法与其他方法在各类别上的F值对比

句子类别	Kim等[8]	A_MQ[10]	本文方法
Background	0.72	0.78	0.80
Intervention	0.16	0.35	0.46
Outcome	0.82	0.86	0.87
Population	0.47	0.51	0.57
Other	0.30	0.84	0.72
Study Design	0.54	0.58	0.47

在循证医学领域，科技文献的结构化表示尚处于探索阶段，本文仅仅是在摘要层次实现了摘要文本的细粒度表示。若要从真正意义上帮助临床医生及研究人员解决从大量科技文献中迅速明确地定位总结出有效临床证据的需求，还需要从全文出发，寻找结构化表示的方案，深入分析科技文献表述临床研究的特征，抽取有用的证据信息，为医生在临床决策和相关研究人员展开进一步研究时提供明确、有效而全面的证据。

参考文献

[1] ABSHER J R. Evidence−Based Medicine [J]. British Medical Journa, 1996, 7(7046):1611.

[2] REJ R. Centre for Evidence−Based Medicine [M]. Springer Netherlands, 2008.

[3] 李幼平. 循证医学[M]. 北京:高等教育出版社, 2015:34−35.

[4] RICHARDSON W S, WILSON M C, NISHIKAWA J, et al. The well−built clinical question: A key to evidence−based decisions [J]. Acp Journal Club, 1995, 123(3): A12.

[5] ELLEN Fineout−Overholt RN, LINDA Johnston RN. Teaching EBP: Asking searchable, answerable clinical questions [J]. Worldviews on Evidence- based Nursing, 2005, 2(3):157−160.

[6] SCHLOSSER R W, O'NEIL−PIROZZI T M, SAMUELSON P A. Problem formulation in evidence−based practice and systematic reviews [J]. Pirozzi, 2006, 33: 5−10.

[7] DAWES M, PLUYE P, SHEA L, et al. The identification of clinically important elements within medical journal abstracts: Patient−Population−Problem, Exposure−Intervention, Comparison, Outcome, Duration and Results (PECODR) [J]. Informatics in Primary Care, 2007, 15(1):9−16.

[8] NAM K S, DAVID M, LAWRENCE C, et al. Automatic classification of sentences to support Evidence Based Medicine [J]. Bmc Bioinformatics, 2011, 12(2):1−10.

[9] DEMNER−FUSHMAN D, LIN J J. Answering clinical questions with knowledgebased and statistical techniques [J]. Computational Linguistics, 2007, 33(1):63−103.

[10] SARKER A. An Approach for automatic multi−label classification of medical sentences [C]. Louhi. 2013.

[11] ARONSON A R. Effective mapping of biomedical text to the UMLS Metathesaurus: The MetaMapprogram[C]// Proceedings/AMIA. 2001.

[12] 刘鸣. 系统评价、Meta: 分析设计与实施方法[M]. 北京:人民卫生出版社, 2011:54−58.

[13] WAARD A D, MAAT H P. Verb form indicates discourse segment type in biological research papers: Experimental evidence [J]. Journal of English for Academic Purposes, 2012, 11(4):357−366.

[14] CORTES C, VAPNIK V. Support−vector networks [M]. Kluwer Academic Publishers, 1995.

[15] XU S, AN X, QIAO X, et al. Multi−task least−squares support vector machines[J]. Multimedia Tools & Applications, 2014, 71(2):699−715.

[16] GIROSI F. An equivalence between sparse approximation and support vector machines[J]. Neural Computation, 1998, 10(6):1455.

[17] AMINI I, MARTINEZ D, MOLLA D. Overview of the ALTA 2012 shared task[J]. Melbourne Australian Language Technology Association, 2013:124−129.

（上接第34页）

[5] QUINTILESIMS INSTITUTE. Outlook for Global Medicines through 2021[EB/OL].[2018−09−10].https: //www.QuintilesIMS.com.

[6] 科学技术部.“十三五”健康产业科技创新专项规划[S].2017.

[7] EWS PROVIDED BY Reportlinker.Top 10 trends in agricultural biologicals market industry: Global forecast to 2022[EB/OL].[2018−09−10].https: //www.reportlinker.com.