面向 PDF 文档的论文元数据混合提取方法

张付志，刘华中，周全强
（燕山大学信息科学与工程学院,秦皇岛 066004）

摘要：针对现有的论文元数据提取方法存在的缺陷与不足，提出一种面向 PDF 文档的论文元数据混合提取方法。首先，根据 PDF 格式论文的特点选取特征，并采用现有数据集对 HMM、SVM 和 CRF 三种统计学习方法进行训练和验证；然后，根据验证结果分别统计三种方法提取各单项元数据的精度值，并借助最大值规则确定每类元数据所采用的提取方法，生成混合提取模型；最后，采用基于时间段统计的方法动态更新提取模型，以保持模型的有效性。实验结果表明，本文提出的方法不仅提高了元数据提取精度，而且具有较强的适应能力。

关键词：元数据提取；统计学习；特征选取；最大值规则；混合提取模型
中图分类号：TP393

Metadata Hybrid Extraction Approach for PDF Document Papers

ZHANG Fuzhi, LIU Huazhong, ZHOU Quanqiang
(School of Information Science and Engineering, YanShan University, Qinhuangdao 066004)

Abstract: Aim at the deficiencies of existing metadata extraction methods, in this paper we propose a hybrid approach to extract metadata from PDF document papers. Firstly, we select features according to the characteristics of PDF format papers, and use existing datasets to train and validate the three statistical learning methods (i.e. HMM、SVM and CRF). Secondly, we calculate each single precision of metadata extraction for the three methods based on the validation results, utilize the maximum rule to identify extraction method for each kind of metadata and generate hybrid extraction model. Finally, we use statistical method based on time period to dynamically update the hybrid extraction model in order to ensure its effectiveness. Experimental results show that the proposed approach not only enhances the precision of metadata extraction, but also has better adaptability.

Key words: metadata extraction; statistical learning; feature selection; maximum rule; hybrid extraction model

0 引言

随着大量开放存取（Open Access，OA）期刊的出现，如何有效组织和管理 OA 期刊论文已成为当前图书馆数字资源建设中的一项重要研究课题。论文的元数据通常包括标题、作者、摘要和关键字等信息。利用元数据对 OA 期刊论文进行组织和管理，可以提高论文检索的准确性和快捷性。因此，如何自动提取 OA 期刊论文的元数据，是数字资源库建设中需要解决的一个关键问题。

论文元数据的自动提取主要有基于规则的方法[1-2]和基于统计的机器学习方法。基于规则的方法需要事先构造提取规则集，利用这些规则从文档中提取元数据。该方法需要预先人工设计提取规则，并且要求规则编制人员对应用领域有深入的了解，如果提取目标发生变化

基金项目：教育部科技发展中心网络时代的科技论文快速共享专项研究资助课题(20101333110013，2011109)
作者简介：张付志(1964-)，男，教授、博士生导师，主要研究方向：智能网络信息处理、网络与信息安全、面向服务计算.刘华中（1982-），男，硕士研究生，主要研究方向：机器学习，智能网络信息处理.周全强（1985-），男，博士研究生，主要研究方向：智能网络信息处理. E-mail: xjzfz@ysu.edu.cn

就会出现规则不适应的情况，因此适应性较差。而基于统计的机器学习方法通过大量的训练样本来训练机器学习算法，经过训练后的算法能够自动处理新数据。例如，Seymore[3]和刘云中等人[4]分别采用隐马尔科夫模型（HMM）实现论文元数据提取，其中各个状态之间需要做独立性假设，没有考虑上下文信息。Ojokoh 等人[5]提出了一种基于三阶隐马科夫模型的引文元数据提取方法，采用三元文法模型有效利用前后状态之间的信息，使提取性能得到提升。周顺先[6]采用基于最大熵的马尔科夫模型提取论文元数据，通过引入状态间转移的条件概率，提高了提取精度。但是，由于采用的是局部模型，只能达到局部最优，并且存在标注偏置问题。Peng 等人[8]提出了一种基于条件随机场（CRF）的论文元数据提取方法，不仅能有效利用上下文特征，而且解决了序列标注偏置问题，提取精度得到提高。Lin 等人[9]提出了一种基于条件随机场的论文元数据提取方法，该方法在提取含有特定参数的文本时效果很好，但在提取作者信息时效果欠佳。Han 等人[11]提出了一种基于支持向量机（SVM）的论文元数据提取方法，该方法能够快速、有效地处理小样本数据，但在处理大规模数据时效率较低。Marinai 等人[12]提出了一种基于神经网络分类器的论文元数据提取方法，该方法能够对会议论文的标题和作者进行有效提取，但是需要大规模的训练样本，并且可靠性较差。张铭等人[13]提出了一种混合元数据提取模型（SVM+BiHMM），通过将支持向量机和二阶隐马尔科夫模型进行融合，元数据提取精确度要优于单一的 HMM 和 SVM 方法。但是，这种串行融合机制仍存在个别元数据提取精度偏低、适应能力差的问题。

针对现有元数据提取方法存在的缺陷和不足，本文提出一种面向 PDF 文档的论文元数据混合提取方法（Metadata Hybrid Extraction Approach for PDF document Papers，MHEAPP）。在 HMM、SVM 和 CRF 三种统计学习方法的基础上，提出了一种面向 PDF 文档的论文元数据混合提取模型。通过对 HMM 模型、SVM 模型和 CRF 模型提取论文元数据结果的评价，利用最大值规则实现元数据提取结果的决策，生成混合提取模型。采用基于时间段统计的方法定期对训练集、验证集和三种提取模型进行更新，以便生成新的混合提取模型。

1 相关知识

1.1 HMM

隐马尔科夫模型（Hidden Markov Model, HMM）[3-5]是一种输出符号序列的统计学习模型，包含观察层和隐藏层。观察层是待识别的观察序列，隐藏层是一个马尔可夫过程，其中每个状态转移都带有转移概率。

设 HMM 具有 n 个状态，由以下五部分组成：状态集合 Q（初始状态 q_1 和结束状态 q_n）、初始状态概率矩阵、状态转移概率矩阵、分散的输出符号词汇表、输出概率矩阵。

HMM 实现过程：首先，由初始状态开始，转移到下一个状态，同时输出一个符号；然后，由该状态转移到另一个状态，转移时输出另一个符号，按照此方式进行，直到转移到结束状态，并且每次状态转移时都会输出一个符号；最后，根据每次转移时输出的符号，得到输出符号序列 $X=x_1x_2\ldots x_l$。在整个过程中，我们能观察到输出符号序列，而状态转移序列不能被观察到。因此，状态转移到哪一个状态，以及转移时输出什么符号，分别由状态转移概率和转移时输出符号概率来决定。

1.2 SVM

支持向量机（Support Vector Machine, SVM）是通过某种事先选择的非线性映射将输入向量映射到一个高维特征空间，在这个空间中构造最优分类超平面。假设有二元分类问题的

训练样本$(x_i,t_i)(i=0,1,\cdots,n)$，其中$x_i$是词$w_i$的特征向量，$t_i\in\{+1,-1\}$为对应的类别标识。SVM 分类器给出输入特征向量$\boldsymbol{x}$的决策函数$f(x)$，其目的是预测未知样本$x$的分类$t$。最优分类函数为[10]：

$$f(x)=sign\left(\sum_{z_i\in SV}a_it_iK(x,z_i)+b\right) \tag{1}$$

式中，a_i是非零系数，z_i属于支持向量，t_i是x对应的类别标注，$K(\bullet)$为核函数。对于非线性问题，仅仅依靠核函数会导致目标空间维数过高，为此引入松弛变量和惩罚因子来解决该问题。

1.3 CRF

条件随机场[7]（Conditional Random Fields, CRF）是一种用于标准和切分有序数据的条件概率模型，融合了最大熵模型和 HMM 的特点，非常适用于自然语言处理领域当中的标注序列化数据的任务。CRF 模型[8]定义了给定文本词序列$\{w_i\}$ (i=1,…,n)，标注序列$\{t_i\}$ (i=1,…,n)的条件概率。其计算公式如下：

$$p(t\mid w)=\frac{1}{Z_0}\exp\left(\sum_{i=1}^{n}\sum_{j=1}^{m}\lambda_jf_j(t_{i-1},t_i,w,i)\right) \tag{2}$$

$$Z_0=\sum_{t}\exp\left(\sum_{i=1}^{n}\sum_{j=1}^{m}\lambda_jf_j(t_{i-1},t_i,w,i)\right) \tag{3}$$

其中，归一化参数Z_o是对于输入观察序列的所有状态序列的标准化因子，表示所有可能序列的得分，使得在输入观察序列上所有可能状态序列的概率之和为 1。而$f_j(t_{i-1},t_i,w,i)$是特征函数，它可以是 0 或者 1，也可以是任意实数，它是对状态转移$t_{i-1}\to t_i$、整个观测序列w以及当前步骤的各方面的一个衡量；λ_j是训练数据对模型进行训练之后，相应特征函数的权重。

2 基于统计学习的论文元数据混合提取方法

2.1 基于统计学习的论文元数据混合提取模型

针对现有的论文元数据提取方法存在的缺陷与不足，本文以 HMM、SVM 和 CRF 三种统计学习方法为基础，设计了一种基于统计学习的论文元数据混合提取模型，如图 1 所示。

该模型主要包括训练、验证、混合提取和模型更新四个功能模块。训练模块首先对 PDF 格式的论文首部进行预处理，根据论文首部的特点进行特征选取，并对选取的特征进行泛化处理；然后借助标注的训练集，训练 HMM、SVM 和 CRF 三个统计学习方法，生成相应的提取模型。验证模块首先借助验证集对三种提取模型进行验证，并计算三种提取模型的单项元数据（例如，标题）提取精度值；然后采用最大值规则判决出提取各单项元数据的最优方法，生成混合提取模型。混合提取模块借助测试集对混合提取模型进行测试，评价其性能。更新模块根据统计测试模块未能正确提取的论文元数据，定期对训练集和验证集进行更新，最终实现混合提取模型的重新决策。

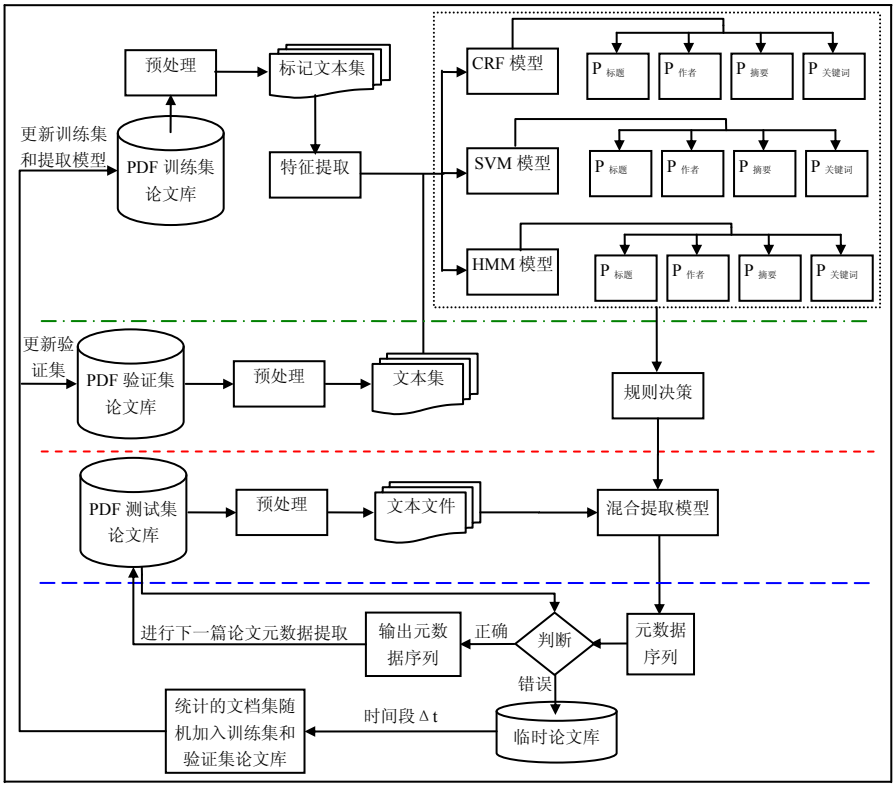

图 1　基于统计学习的论文元数据混合提取模型

Fig.1 Metadata hybrid extraction model based on statistical learning for PDF document papers

2.2　特征选取

为了便于特征选取，需要对 PDF 格式的论文进行预处理。首先，利用开源工具 PDFBOX 将 PDF 格式论文的首页转换为文本格式；然后，采用正则表达式去掉文本中的冗余信息（例如，出版日期、期刊名称、刊号、作者详细信息、url 和 Email 等），生成目标信息。在此基础上，采用文本分块技术将目标信息划分为块序列 block1、block2、block3、block4，并且以“1 Introduction”、“1 引言”等之类的信息作为最后一个分块的结束。

针对 PDF 格式论文元数据的提取问题，在选取特征时需要考虑以下几个方面。

(1)局部文本特征：文本输入序列局部所特有的字符以及与排版相关的特征。如作者名称简写形式，一行的开始和结束等。

(2)外部词典特征：分析的对象是一些特殊的字或者词，如作者名称包括 firstname、lastname、femalename 和 malename 词典库，论文中出现的高频词（如 abstract 和 keywords 等）作为高频特征词。

(3)状态转移特征：特征函数 $f_j(t_{i-1},t_i,w,i)$ 整合输入数据序列 w 和状态转移 $t_{i-1}\to t_i$ 的特征。例如 t_{i-1} 的状态是标题，t_i 的状态是作者，并且 w_i 是出现在人名数据库 firstname 字典库里的词，则特征函数值取 1。本项特征仅适合 CRF 模型。

2.3　基于混合提取模型的论文元数据提取

针对各种提取方法在实现论文元数据提取时出现单项元数据精度值大小不一的问题，本文采用最大值判决规则实现元数据提取结果的融合，最大程度地发挥各种提取方法的优势。

2.3.1　基于 HMM 的论文元数据提取

应用 HMM 模型进行论文头部元数据提取时，隐藏层是由标题、作者等元数据组成的有

限状态机，观察层是待提取的论文头部文本序列。具体实现过程如下：

首先，对训练集文本块进行预处理，分别对块序列进行状态标记，应用 ML(Maximum Likelihood)算法得出 HMM 模型参数，分别计算初始状态概率、转移状态概率和状态输出概率。其中，状态转移概率以块为单位，状态输出概率以单词为单位。

然后，将待提取的论文经预处理生成的块序列作为模型输入，同时计算各块的输出概率，采用 Viterbi 算法动态规划预测出最大概率的状态序列，该状态序列就是本文要提取的元数据序列。

2.3.2 基于 SVM 的论文元数据提取

如果把每项元数据视为一类，元数据的提取就转化为对每个文档块进行分类的问题。

首先，确定要提取的元数据（标题、作者、摘要和关键词），对论文头部进行分块处理，并在块内进行特征选取，建立每行的特征向量。

然后，对验证集进行分类，根据分类结果，借助某块的前后 2 块来建立该块的新特征，对原特征向量进行修改，并用迭代的 SVM 分类器对修改后的特征向量重新进行分类。

最后，对文本块序列应用一对多的 SVM 多分类器，实现块分类，每类就是相应的元数据类别。

2.3.3 基于 CRF 的论文元数据提取

根据论文头部的特点选取合适的特征后，利用 CRF 方法提取论文元数据的操作为：

首先，对经预处理的训练文本进行分块，并对其进行状态标记，采用 L-BFGS 算法从标注好的训练集学习 CRF 模型的参数。

然后，对输入的论文头部进行分块预处理，生成的块序列为 w，将其输入由公式(2)定义的 CRF 概率模型，分别计算各块的输出概率和状态转移概率，利用公式（4）求出最可能的状态序列：

$$t^* = \arg\max_t P(t / w) \qquad (4)$$

其中，t^* 作为 w 的状态输出，也即是各元数据状态序列。

2.3.4 混合提取模型的融合策略

为了发挥每种提取方法的优势，以 HMM、SVM 和 CRF 三种统计学习方法提取论文元数据的结果为依据，采取三种方法融合的策略实现论文元数据的提取。考虑到每种方法在提取同一元数据时出现提取精度值不一致的问题，根据每种方法的提取结果，我们借助最大值判决函数确定提取某元数据精度值最高的方法，并将其作为提取该单项元数据的提取方法。具体实现如下：

假设提取模型集合 X=$\{M_i|1\leqslant i\leqslant I\}$（$I$ 为提取模型个数），元数据集合 Y=$\{S_c|1\leqslant c\leqslant C\}$（$C$ 表示元数据类别个数），M_i 表示 X 中的第 i 项，S_c 表示 Y 中的第 c 项；则判决第 c 项元数据 S_c 的最大提取精度的计算公式如下：

$$P_{S_c} = Max(P_{S_cM_1}, P_{S_cM_2}, ..., P_{S_cM_i}, ..., P_{S_cM_I}) \qquad (5)$$

根据公式（5）可以得出提取 S_c 所采用的最优方法，采用同样判决规则，我们可以得出提取其它类别的元数据所采用的最优方法。在进行论文元数据提取时，集成各个最优方法提取相应类别的元数据，生成混合提取模型。

2.4 混合提取模型的更新策略

随着时间的推移，会出现越来越多论文的元数据不能被正确提取，也即是混合提取模型不能根据环境的改变进行更新。为了解决这一问题，本文采用基于时间段统计的方法定期更新混合提取模型。

首先，利用测试集对混合提取模型进行测试，根据提取的论文元数据和原始的论文元数据比对，判断提取的论文元数据是否正确：若提取正确，则输出元数据序列，并进行下一篇论文元数据的提取；若提取错误，则将原始的论文从测试集论文库中取出，并存入临时文档库。

然后，设定时间段 Δt，每经过一个 Δt 对临时文档库里的论文集进行统计，并将统计的论文集随机分散到原训练集和原验证集的论文库中，同时清空临时文档库，以存取下一个时间段未能正确提取的论文。

最后，选取 Δt 内临时论文库中论文的特征并将其存入原特征库中，并借助已更新的训练集对各个提取方法进行再训练，生成相应的新提取模型，同时利用已更新的验证集对它们进行评价，采用最大值判决规则实现提取结果的再次决策，得出混合提取方法中提取各元数据的相应提取方法，重新生成混合提取模型。

3 实验结果与分析

3.1 实验数据

本文实验数据集通过下载各 OA 期刊的 6000 篇 PDF 格式论文产生。首先从经过预处理的 6000 篇论文中排除 500 篇与其它论文格式截然不同的论文，然后从剩余的 5500 篇论文中随机选取 2500 篇论文构成数据集 F，其余的 3000 篇和排除的 500 篇论文构成数据集 S。提取的元数据包括：Title、Author、Abstract 和 Keywords。

3.2 评价指标

为了评价论文元数据提取方法的性能，采用各类元数据的提取精度值 *P*（*Precision*）作为评价指标。其计算公式如下：

$$P=\frac{\text{正确提取该类别元数据个数}}{\text{某类别元数据总数}}\times 100\% \tag{6}$$

为了评价论文元数据提取方法的适应性，采用整体提取精度值 *WP*（*Whole Precision*）作为评价指标，其计算公式如下：

$$WP=\frac{\text{正确提取所有类别元数据个数}}{\text{元数据总数}}\times 100\% \tag{7}$$

同时，为了保证输出结果精度值的有效性，实验进行 *J* 折交叉验证，将数据集分成 *J* 份，轮流将其中 *J*-1 份作为训练集，1 份作为测试集，进行实验。每次实验都会得出相应的提取精度值，取 *J* 次平均值 $\overline{P}$ 。其计算公式如下所示：

$$\overline{P}=\frac{\sum_{j=1}^{J}P_j}{J} \tag{8}$$

为了进一步降低误差，又进行 *H* 次 *J* 折交叉验证，并取平均值 P_{avg}，其计算公式如下所示：

$$P_{avg}=\frac{\sum_{h=1}^{H}\overline{P}_h}{H} \tag{9}$$

实验参数设置：H 和 J 值均取 10。

3.3 各种提取方法的精度比较

为了评价各类元数据的提取精度，利用数据集 F 将本文提出的 MHEAPP 分别与以下几种方法进行对比：文献[4]的 HMM、文献[11]的 SVM、文献[8]的 CRF、文献[13]的 SVM+BiHMM。

我们从 F 数据集随机选取 2000 篇，分别对以上五种方法进行 H 次 J 折交叉验证。其中，在对 MHEAPP 实验的过程中，我们首先从 F 数据集随机选取 2000 篇，分别对本文采用的 HMM、SVM 和 CRF 进行 H 次 J 折交叉验证；然后，根据统计的提取结果，采用最大值判决规则，得出标题、作者、摘要和关键词的相应提取方法分别为 SVM、HMM、CRF 和 CRF，组成混合提取方法；最后再从 F 数据集随机选取 2000 篇，再对混合提取方法进行 H 次 J 折交叉验证。表 1 给出了五种元数据提取方法所得到的单项元数据提取精度，根据表 1 中的数据生成的元数据提取精度对比图如图 2 所示。

表 1 五种方法对各类元数据的提取精度（%）

Tab. 1 Precision of the five methods for each type of metadata (%)

Metadata	HMM	SVM	CRF	SVM+BiHMM	MHEAPP
Title	90.6	93.5	93.2	93.6	93.4
Author	93.2	91.6	92.1	92.5	93.3
Abstract	92.9	94.8	95.7	95.2	96.0
Keywords	91.4	92.3	93.9	92.2	94.1

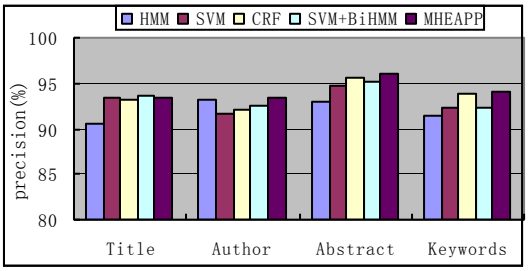

图 2 五种方法对各类元数据提取精度的对比

Fig.2 Comparison of the five methods' precision for each type of metadata

从表 1 和图 2 可以看出，SVM+BiHMM 方法相对于 SVM 和 HMM 两种方法的提取性能确实有所提升，特别是 Abstract 的精度得到很大提高，但是 Author 和 Keywords 的精度相对较低，因此整体提取性能提高不太明显。MHEAPP 通过融合 HMM、SVM 和 CRF 三种方法的提取结果，各单项元数据的精度均达到了最优，整体提取性能均优于三种方法。与 SVM+BiHMM 方法相比，MHEAPP 方法的单项元数据提取精度，除了 Title 降低 0.2%外，其余部分的精度值均比 SVM+BiHMM 方法的高。

3.4 MHEAPP 的适应性评价

为了验证 MHEAPP 的适应性，我们采用数据集 S，把它随机分成 10 等份，并将实现每一等份的论文元数据提取假设为一个时间段 Δt，共分成 10 个时间段。本次实验首先将 MHEAPP 分成两组：其中一组是不考虑对混合提取模型进行更新，称为未带更新的 MHEAPP；另一组是对混合提取模型进行定期更新，称为带更新的 MHEAPP。然后将二者与以下几种方法进行了对比：文献[4]的 HMM、文献[11]的 SVM、文献[8]的 CRF、文献[13]

的 SVM+BiHMM。实验评价标准是论文元数据的整体提取精度 *WP*，实验结果如图 3 所示。

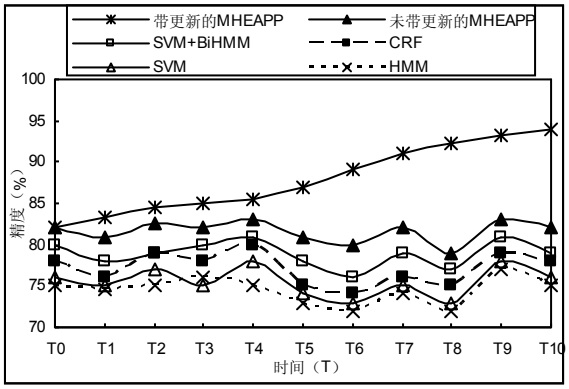

图 3 六种方法的整体提取精度对比

Fig.3 Comparison of the six methods' whole precision

在元数据提取过程中，实验数据被随机加入了具有新特征的论文，从图 3 可以看出，由于不带更新的 MHEAPP、HMM、SVM、CRF 和 SVM+BiHMM 均不具有更新提取模型的功能，所以这五种方法提取元数据的精度值都一直相对较低。而带更新的 MHEAPP 在整个提取过程中，元数据提取精度值逐渐增大：在每个时间段的结束，将临时文档库中的论文随机分成两部分别存入原训练集和验证集论文库中，同时把临时文档库中的论文新特征加入原特征库中，借助新的训练集和验证集实现混合提取模型的更新，进而使其逐渐恢复到高性能状态。这说明，对混合提取模型定期更新的 MHEAPP 能够处理具有新特征的论文，适应性强。

4 结论

针对现有论文元数据提取方法提取精度不高且适应能力较差的问题，本文提出一种面向 PDF 文档的论文元数据混合提取方法。该方法借助训练集对集成的三种提取方法进行建模，并利用验证集对各个模型进行验证评价，采用最大值规则实现论文元数据提取结果的决策，生成混合提取模型。另外，为了保证混合提取模型的有效性，采用基于时间段统计机制对混合提取型进行定期更新。同现有的元数据提取方法相比，本文提出的方法不仅提高了元数据精度，而且具有较强的适应能力。

[参考文献] (References)

[1] Councill I G, Giles C L, Iorio E D, Gori M, Maggini M, Pucci A. Towards Next Generation CiteSeer: A Flexible Architecture for Digital Library Deployment[A]. 10th European Conference on Digital Libraries[C]. Berlin：Springer-Verlag，2006. 111-122.

[2] Flynn P, Zhou L, Maly K, Zeil S, Zubair M. Automated Template-Based Metadata Extraction Architecture[A]. In Proceedings of ICADL[C]. Berlin: Springer-Verlag, 2007. 327-336.

[3] Seymore K, McCallum A, Rosenfeld R. Learning Hidden Markov Model Structure for Information Extraction[A]. In AAAI 99 Workshop on Machine Learning for Information Extraction[C]. Palo Alto: AAAI Press,1999. 37-42.

[4] 刘云中，林亚平，陈治平. 基于隐马尔可夫模型的文本信息提取[J]. 系统仿真学报，2004，16（3）：507-510.

[5] Ojokoh B, Zhang M, Tang J. A Trigram Hidden Markov Model for Metadata Extraction from Heterogeneous References[J]. Information Sciences, 2011, 181(9):1538-1551.

[6] 周顺先. 文本信息提取模型及算法研究[D]. 长沙：湖南大学，2007.

[7] Lafferty J, McCallum A, Pereira F. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data[A]. In Proceedings of the Eighteenth International Conference on Machine Learning[C]. San Francisco: Morgan Kaufmann, 2001. 282-289.

[8] Peng F, McCallum A. Accurate Information Extraction from Research Papers using Conditional Random Fields[A]. In Proceedings of Human Language Technology Conference and North American Chapter of the Association for Computational Linguistics[C]. New York: ACM Press, 2004: 329-336.

[9] Lin S, Ng J P, Pradhan S, Shah J, Pietrobon R, Kan M Y. Extracting Formulaic and Free Text Clinical

Research Articles Metadata Using Conditional Random Fields[A]. In Proceedings of the NAACL HLT 2010 Second Louhi Workshop on Text and Data Mining of Health Documents[C]. Los Angeles: Association for Computational Linguistics Press, 2010. 90-95.

[10] Campbell C, Ying Y. Learning with Support Vector Machines[M]. San Rafael: Morgan and Claypool Publishers, 2011.

[11] Han H, Giles C L, Manavoglu E, Zha H. Automatic Document Metadata Extraction using Support Vector Machines[A]. In Proceedings of the 2003 Joint Conference on Digital Libraries[C]. Washington: IEEE Computer Society, 2003. 37-48.

[12] Marinai S. Metadata Extraction from PDF Papers for Digital Library Ingest[A]. In Proceedings of the 2009 10th International conference on Document Analysis and Recognition[C]. Washington: IEEE Computer Society, 2009. 251-255.

[13] 张铭，银平，邓志鸿，杨冬青. SVM+BiHMM：基于统计方法的元数据提取混合模型[J]. 软件学报，2008，19（2）：358-368.