Digital Library Forum 数字图书馆论坛
http: // www.dlf.net.cn

基于条件随机场的专利摘要信息抽取研究*

□ 黄绍杉 乔晓东 桂婕 李鹏 / 中国科学技术信息研究所 北京 100038

摘要：专利摘要是重要的情报分析数据来源，但其自然语言文本的特征，使得专利摘要的自动内容抽取具有较大难度。文章利用亚洲语言信息检索测评会议（NACSIS Test Collections for IR，NTCIR）提供的英文专利文摘测试语料，采用文本信息抽取统计方法中的条件随机场模型，通过提取并添加有效的特征，有针对性地抽取专利摘要中表示技术和功效内容的信息，为专利的技术功效矩阵分析提供可机器自动抽取的强大支持。

关键词：专利摘要，信息抽取，条件随机场
DOI：10.3772/j.issn.1673—2286.2010.09.003

1 引言

专利是人类发明创造的智力活动和法律活动的结合和交叉，是人们依据国家法律，对自己的智力活动所获得的成果谋取权利保护的过程[1]。专利历来被视为一种科技发展的重要参考坐标，主要原因在于其作为一种特殊的文献类型，专利信息涵括了工业产权情报、技术情报、商业经济情报，是一个具有战略研究意义的情报源，被广泛深入地应用于科技发展态势及分布、技术前沿的分析研究，为国家的科技战略层决策、企业科技创新提供参考与分析依据，可以给广大科研人员和工程技术工作人员提供方法与建议，成为一股重要的推动力[2]。

专利信息分析就是从专利文献中采集专利信息，通过科学的方法对专利信息进行加工、整理和分析，最终形成专利情报和谋略的一类科学劳动的集合[2]。目前，专利分析中普遍使用的分析指标多针对专利的著录项信息，如专利所在国别、专利发明人、专利申请人、专利分类号、专利申请日等，而对专利技术内容的挖掘需要对专利的文本内容进行处理与分析，其利用程度不够深入。专利摘要中记载的大量技术内容特征（如反映技术的改进、换代、新颖性、功能用途、关键技术要素等重要细节性内容）因受技术限制、人工成本和时间的限制，为此一直没得到很好的利用，限制了专利分析方法的进一步发展。本研究尝试利用条件随机场技术进行专利摘要信息的抽取。

2 专利摘要的信息抽取技术

信息抽取可以准确地抽取出用户所需要的具体信息，而不只是相关的文档，是指从文本中抽取用户感兴趣的事件、实体和关系，被抽取出来的信息以结构化的形式进行描述，然后存储在数据库中，为用户进一步使用。基于信息抽取技术构建的信息抽取系统其处理的对象为自然语言文本尤其是非结构化的文本信息，从广义角度而言，本文所指的信息抽取技术处理对象还包括了语音、图像、视频等其他的媒体类型的数据，即本文所关注的信息抽取技术为狭义上的信息抽取研究，局限于对自然语言文本的信息抽取技术[3]。

如上文所提到的，作为自然语言文本的专利文献摘要部分，人工处理已经远远不能满足当前专利信息分析研究工作的要求，因其需要把信息抽取技术应用到专利摘要的内容特征分析上，从专利摘要中抽取出其中的技术关键性单词、短语、句子等。信息抽取

* 本文得到国家科技部“十一五”科技支撑计划（项目编号：2006BAH03B03）、中国科学技术信息研究所重点工作项目（项目编号：2009KP01—7—1）、中国科学技术信息研究所2009年度预研基金项目（项目编号：YY—200906）等项目的资助。

技术主要涉及两种方法：一是知识工程方法（Knowledge Engineering Approach）；二是自动训练方法（Automatic Training Approach）。知识工程方法主要靠手工编制规则使系统能处理特定知识领域的信息抽取问题。这种方法要求编制规则的知识工程师对该领域知识有深入的了解，且规则制定的过程比较耗时耗力。自动训练方法系统主要通过学习已经标记好的语料库获取规则，并且经训练后的系统能自动学习处理新的文本[4]。

本文的专利摘要信息抽取研究的目的是期望通过对大量科技论文或者专利文献的自动化抽取处理，生成技术功效矩阵表。由矩阵表中的各区域的密度分布，可看出技术密集区、地雷禁区、尚未被开发区域或有益可图的技术领域，以为企业的技术创新管理提供支持。

表1 技术功效矩阵表[5]

	Effect 1	Effect 2	Effect 3
Technology 1	[AAA 1993] [USPat.XX/XXXX]		[BBB 2002]
Technology 2	[CCC 2000]		
Technology 3		[USPat.YY/YYYY]	[US Pat.ZZ/ZZZZ] [USP WW/WWW]

3 专利摘要信息抽取流程

本文根据条件随机场（CRFs，参见3.1小节）的运行要求和系统目标设计了针对专利摘要信息抽取的流程，主要包括数据预处理、特征分析与选择、模板构建、模型训练和测试几个部分，数据预处理过程把基础语料处理成符合条件随机场模型输入的格式要求，这一过程需要把可利用的语料特征加入到模型中去，并结合相匹配的特征模板，执行模型训练的过程，也就是一个机器学习的过程，然后使用训练好的CRFs模型对预留的测试语料进行测试，最终形成一个序列标注结果，并根据标注结果设定抽取规则形成最后的抽取结果文档。最后，使用设定的评估标准进行定性评估，分析抽取效果。

本文的专利摘要信息抽取研究流程如图1所示。

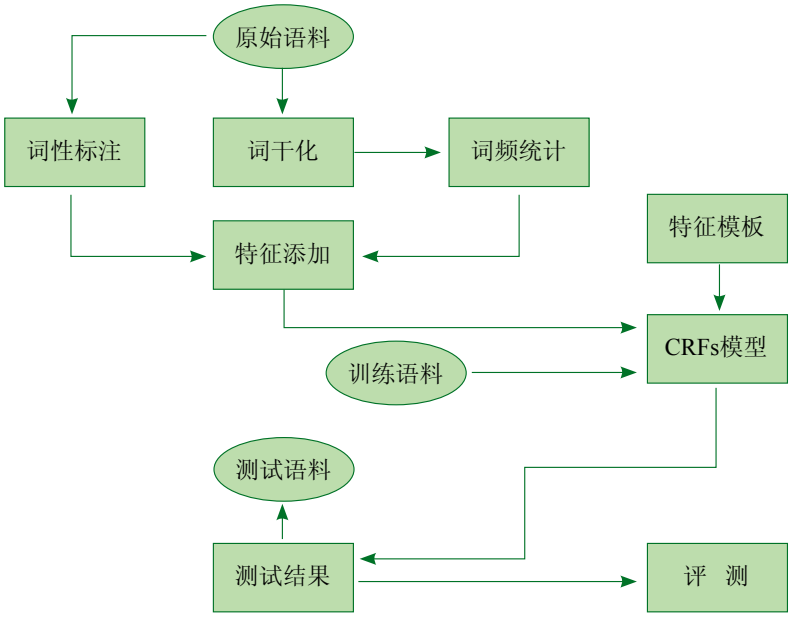

图1 专利摘要信息抽取研究流程

3.1 条件随机场

条件随机场（Conditional Random Fields，CRFs）是信息抽取技术中常用的概率统计模型，是一种无向图模型，可用于最大化条件概率。CRFs最早由Lafferty等人[4]于2001年提出，其思想主要来源于最大熵模型（Max entropy）。可以把CRFs看成是一个无向图模型或马尔可夫随机场，它是一种用来标记和切分序列化数据的统计框架模型。目前，CRFs在解决英文浅层分析、英文命名实体识别等问题时已经取得了良好的效果。

CRFs是一种无向图模型，假设X、Y分别表示需要标记的观察序列和相对应的标记序列的联合分布随机变量，那么CRFs(X，Y)就是一个以观察序列X为条件的无向图模型。定义G=(X，Y，E)为一个无向图，Y={Y_v，U∈V}，即V中的每个节点对应于一个随机变量所表示的标记序列的元素h。如果每个随机变量Y_v对于G遵守马尔可

夫属性，那么(X，Y)就构成一个CRFs，而且在给定X和所有其他随机变量Y(u|u≠v,{u,v}∈V)的条件下，随机变量Yv的概率为：

$P(Y_v|X,Y_u,u\neq v,\{u,v\}\notin V)=P(Y_v|X,Y_u,\{u,v\}\notin E)$

在图形模型中的各输出结点被连接成一条线性链的特殊情形下，CRFs假设在各个输出结点之间存在一阶马尔可夫独立性，二阶或更高阶的模型可类似扩展。若让$O=(O_1,O_2,\cdots,O_T)$表示被观察的输入数据序列，让$S=(S_1,S_2,\cdots,S_T)$表示一个状态序列。在给定一个输入序列的情况下，线性链的CRFs定义状态序列的条件概率为：

$$P_A(S|O)=\frac{1}{Z_0}\exp\left(\sum_{t=1}^{T}\sum_{k=1}^{K}\lambda_k f_k(S_{t-1},S_t,O,t)\right)$$

在模型的算法复杂度、特征选取兼容性、数学理论基础等方面，条件随机场CRFs和另一常用的隐马尔科夫HMM相比都有优势，优势主要在于以下几点[5]：

（1）能够在同一个模型中无限制集成不同特征，特别是可加入远距离约束，更能揭示语言学特征；

（2）CRFs采用联合条件概率P(T/W)建模，避免了HMM的独立性假设和二元假设，具有更合理的数学推导；

（3）CRFs保留了HMM中的之前标记的状态对当前状态标记的影响，使特征的选择更为合理；

（4）CRFs采用无向图模型，是对整个标记序列求解联合概率，在整个序列范围内归一化，较HMM具有更合理的数学理论基础，同时也避免了因求解局部观察值概率所带来的标记偏置问题。

在实证研究方面，研究人员已经进行了角色识别、科技术语抽取、地名识别等应用，证明了CRFs的高效性，因此，笔者在本文中采用CRFs模型作为信息抽取的统计工具。

3.2 数据来源及其特征选取

本文的专利数据来源于日本NTCIR第八次会议的专利测评比赛测试数据集，内容为美国专利及商标局英文专利文摘，文摘中的信息抽取对象已经进行了手工标注，具体标签结构和含义如表2所示。

数据标注范例如下：

表2 标签及其含义[6]

标签名称	标签含义
<TECHNOLOGY>	在每个研究或发明中使用到的算法、工具、材料或数据
<EFFECT>	功效标签中包含一对“属性”和“价值”标记
<ATTRIBUTE>	技术的功效可以被一对“属性”和“价值”所表示
<VALUE>	被手工分配给每篇论文/专利的IPC号码

<TOPIC>
<TOPIC-ID>6</TOPIC-ID>
<IPC-LIST><IPC>H04N_1_40</IPC><IPC>B41J_2_525</IPC><IPC>G03G_15_01</IPC><IPC>G06F_3_12</IPC><IPC>G06F_15_62</IPC><IPC>H04N_1_46</IPC></IPC-LIST>
<TEXT>
<TITLE>Method and apparatus for recording color images in both interlaced and non-interlaced modes</TITLE>
<ABSTRACT>An image recording apparatus having plural recording methods and permitting selection of a recording method in accordance with a desired level of record results during recording of image data to produce a difference in color between the record results due to changes in the recording methods when an image to be recorded is color data. The invention includes <TECHNOLOGY>plural color conversion </TECHNOLOGY> corresponding to a plurality of recording methods so as to perform color conversion suitable for the selected recording <EFFECT><VALUE>prevent</VALUE><ATTRIBUTE>differences in the color of the record results</ATTRIBUTE></EFFECT> from occurring.</ABSTRACT>
</TEXT>
</TOPIC>

上述专利文摘文本中所需提取的字段信息有如下特点：

（1）技术（TECHNOLOGY）名词性短语或者包含名词性短语

的句子，句子中的其他成分是对名词性短语的修饰、解释或者限定。大多数情况特征为名词性短语后面跟介词for连接其他成分，少数以having和capable of等词语连接。

（2）价值（VALUE）大多数为表示对现有状态改变的动词，如improve，对变化控制性的动词如prevent，表示改变效果的形容词如good，带形容词修饰的名词如highly accurate、higher quality，相对应动词的名词形式如prevention。

（3）属性（ATTRIBUTE）为单个名词或者名词性短语，为摘要中提出的技术和发明造成的对原有状态改变的作用对象，长度从一个单词到多个词语修饰的名次性词组，没有复杂长句子。属性（ATTRIBUTE）和价值（VALUE）一般出现在一个独立句子中，位置比较接近。

（4）技术字段因多数为名词性结构，多有冠词起引领作用，价值字段因专利文献写作的规范化，其中有大量的高频词出现。

综上分析，本文选用词性特征、位置特征、冠词特征、频次特征作为特征空间的四个元素。

3.3 数据预处理

在专利摘要信息抽取前需要对语料文本进行预处理，本文把特征空间的四个特征以CRFs所能识别的格式进行添加，具体步骤如下：

第一步：对专利文本进行自动词性标注

词性标注的任务就是根据一个词在某个特定句子中的上下文，为这个词标注正确的词性[7]。本文使用了东京大学计算机科学系TSUJII实验室开发的基于最大熵模型的英文词性标注工具POStagger-1.0[8]，标注速度达到每秒2400，精确度为97.1%。本文利用Java程序模块去除文本中的自带标签，然后使用POStagger-1.0进行词性的自动标注。

第二步：词性特征标示列和词单列的格式转化

词性特征是本文使用的第一种特征，而上面形式的格式是不能直接为CRFs模型工具所处理的，将标注好的文档用程序模块转换成纵列的形式，即每一个词成为第一纵列，后面对应为标注好的词性标示，词性标示表示作为第二纵列。

第三步：词位置属性标签的添加

本文处理的原始文本中加入要抽取的各种字段的文本位置特征，利用已经标识好的标签，标明每个词的位置含义，本文共用七个标签来标示其位置特征，比如用B-TEC表示技术词第一个词，B-ATT表示属性词第一个词，其余类似。

第四步：把两种特征标示字符列进行合并处理

把第二步和第三步实现的两个纵列整合成规范格式，每一行为一个token，各列之间用空格或制表格间隔。一个token的序列构成一个sentence并用一个空行间隔，各列之间用空格或制表格间隔。

上述内容是词性和位置特征添加过程的格式转化步骤，冠词特征和词频特征的添加步骤与此类似，一个主要不同之处是词频特征添加前的词频统计，为了保证名词单复数、各种时态对频次统计结果的噪音干扰，要进行一步词干化的处理，本文使用在线处理模式的snowball词干化工具[9]。本文最终所形成的文档格式样例如下：

表3 输入文档格式

词	词性特征	冠词特征	词频特征	位置特征
The	DT	1	2848	S
Facsimile	NN	0	20	S
Machine	NN	0	23	S
Has	VBZ	0	73	S
a	DT	1	1801	B-TEC

3.4 特征模板选用

本文在具体流程设计中，按照CRFs的要求设计了相对应的特征模板，模板是对上下文环境中的特定位置和特定信息的考虑，反映了所要考虑的语言现象的选取标准，也可以理解为它指导和限定了机器学习过程的空间范围。

特征模板文件中的每一行代表一个template。每一个template中，专门的宏%x[row,col]用于确定输入数据中的一个token，row用于确定与当前的token的相对行数，col用于确定绝对行数[10]。

有两种类型的模板文件，类型可由第一个字符来区分，第一种是Unigram template，第一个字符是U，当给出一个模板"U01:%x[0,1]"，CRFs会自动地

生成一个特征函数集合(func1 … funcN)。

另一种是Bigram template。第一个字符是B，这个模板用于描述bigram features。使用这个模板，系统将自动产生当前输出token与前一个输出token(bigram)的组合。产生的可区分的特征的总数是L*L*N，其中L是输出类别数，N是这个模板产生的unique features数[10]。当类别数很大的时候，这种类型会产生许多可区分的特征，这将会导致训练和测试的效率都很低。

根据本文添加的四个特征，编写了相对应的特征模板文件template，为防止运行效率过低和拟合现象的出现，将“观测窗口”（N值）设定为5，采用二元结构，在CRFs的模型训练程序初始化阶段，由训练模块open template方法负责将上面定义好的特征模板文件读取出来，依据模版的类型加载到其中的二元模板容器bigram template中，作为从训练语料中提取特征的规则和标准。模板文件片断如下：

```
# Unigram
U000:%x[-2,0]
U001:%x[-1,0]
U002:%x[0,0]
U003:%x[1,0]
U004:%x[2,0]
```

4 实验结果与分析评价

具体实验过程中，本文采用了所有语料中的251条作为训练文档集，50条作为测试文档集。

在对抽取性能进行评估时，采用常用的3个评测指标，准确率(P)、召回率(R)、综合指标F值(F)，P和R的计算公式如下[10]：

$$P=\frac{\text{准确抽取的信息条数}}{\text{抽取出的所有信息条数}}$$

$$R=\frac{\text{准确抽取的信息条数}}{\text{应该被抽取出的信息条数}}$$

实际评估时，需要综合考虑P和 R，采用综合评价指标F值进行评价，计算公式如下：

$$F=\frac{(\beta^2+1)PR}{\beta^2P+R}$$

这里，β表示P和R的权重设置，通常设置为1、2、1/2，本文中将β设置为1，即两者重要性相同。

表4所列为本文实验得到的结果提交到NTCIR-8评测会议进行预评审而返回的结果。

表4 NTCIR−8测评结果

	Recall	Precision	F-Measure
Technology	0.165 (34 / 206)	0.351 (34 / 97)	0.224
Value	0.143 (4 / 28)	0.667 (4 / 6)	0.235
Attribute	0.030 (1 / 33)	0.500 (1 / 2)	0.057
Averge(均值)	0.159 (44 / 276)	0.393 (44 / 112)	0.227

由表4的测评结果所示，本文中提交的参赛结果准确率较好，召回率比较低，通过分析，导致各项指标不是很理想的原因如下：

（1）训练文档语料规模较小，限制了用CRFs模型的机器学习效果。用于训练的文档集大小，对模型的抽取效果有比较重要的影响。

（2）训练语料中的抽取对象比较复杂，有短语，也有部分短语带修饰成分的结构，还有少数句子，进一步限制统计特征的明显性。

5 总结

专利摘要信息中提供的技术关键词、功效类词语信息是专利信息分析工作可以利用的重要信息，笔者试图通过计算机的自动处理将这些关键信息自动提取出来，以有助于最大化降低专利分析的人工成本。本文通过对NTCIR-8会议所提供专利语料的特点和识别难点的详细分析，利用信息抽取统计方法中目前应用比较广泛的条件随机场模型，设计了包括数据预处理、特征添加、模型训练与测试等模块的信息抽取系统，并进行了技术功效词抽取的实证研究，取得了较好的抽取结果。

鉴于抽取对象的文本语法结构比较复杂，单纯地利用统计方法其抽取效果提升空间有限，下一步可以通过编订规则弥补统计方法抽取结果中的召回率准确率不高的不足。利用自然语言处理技术应用到专利信息的自动化处理上，已经成为研究的热点，目前出现的信息抽取系统多受限于某一具体领域，如何提高系统的兼容性和扩建性有待研究人员的进一步探讨和研究。

参考文献

[1] 李保力,陈玉忠,俞士汶. 信息抽取研究综述[J]. 计算机工程与应用,2003,39(10):1-2.
[2] 彭爱东. 企业专利情报信息研究[D]. 南京大学,2000.6.
[3] 邓尚民,孙玉伟. 信息抽取系统的研究现状[J]. 现代图书情报技术,2006(3):54-58.
[4] 刘开瑛,郭炳炎. 自然语言处理[M]. 北京:科学出版社,1991.
[5] SETTLES B．Biomedical Named Entity Recognition Using Conditional Random Fields and Rich Feature Sets [C]// Proceedings of the International Joint Workshop on Normal Language Processing in Biomedicine and its Application(NLPBA), Geneva, Switzerland, 2004:103-107.
[6] The 8th NTCIR Workshop [EB/OL]. [2010-08-10]. http://research.nii.ac.jp/ntcir/ntcir-ws8/ws-en.html.
[7] 叶娜. 面向信息抽取的文本预处理和规则自动学习技术研究[D]. 东北大学,2004.
[8] postagger.1.0 [CP/OL]. [2010-08-10]. http://www-tsujii.is.s.u-tokyo.ac.jp/~tsuruoka/postagger/.
[9] snowball [CP/OL]. [2010-08-10]. http://snowball.tartarus.org/demo.php.
[10] 佘丰. 专利摘要的信息抽取研究[D]. 北京理工大学, 2006.

作者简介

黄绍杉(1983-)，中国科学技术信息研究所信息技术支持中心在读硕士研究生，研究方向：专利知识抽取。通讯地址：北京市复兴路15号信息技术支持中心 100038。 E-mail: hsshss1983@163.com
乔晓东(1965-)，硕士，研究员，研究方向：信息服务和信息资源管理。通讯地址同上。E-mail: qiaox@istic.ac.cn
桂婕(1976-)，博士，助理研究员。研究方向：专利分析和科技创新管理。通讯地址同上。E-mail: guij@istic.ac.cn
李鹏(1979-)，硕士，助理研究员。研究方向：智能信息处理。通讯地址同上。E-mail: lipeng_cn@istic.ac.cn

Information Extraction of Patent Summary Based on Conditional Random Fields

Huang Shaoshan, Qiao Xiaodong, Gui Jie, Li Peng / Institute of Scientific and Technical Information of China, Beijing, 100038

Abstract: Patent summary is an important data source of intelligence analysis. However, the characteristics of natural language text make automatic content extraction more difficult. In this paper, we use the testing data provided by NTCIR-8 and the conditional random fields model which is one of the information extraction statistical methods, by extracting and adding effective features, we extract the technology and efficacy information, to provide the machine automatical extraction for technical and efficiency matrix analysis of patent.

Keywords: Patent summary, Information extraction, CRFs

（收稿日期：2010-08-15）

业界动态

维基百科创始人：报纸书籍的寿命还很长

据国外媒体报道，维基百科创始人吉姆·威尔士(Jimmy Wales)日前在接受采访时表示，移动互联网的普及为传统报纸提供了新的营收模式。报纸和书籍的寿命还很长，远没有人们想象得那么糟糕。新闻服务供应商将获得新的盈利途径，即对数字内容进行收费。

威尔士说：“iPad或Kindle的‘应用’模式为报纸提供了新的机会。如果可以通过iPad阅读，价格又合理，相信微支付模式会给人们带来一种购买冲动。”

目前，无线支付系统尚未普及。但威尔士认为，这两种模式均十分有效，但注册模式可能更受欢迎。威尔士说：“我不会从钱包中掏出信用卡，太麻烦了。但如果仅仅点击几下就可以获得所需内容，我愿意为此支付一些费用，因为很值得。”

来源：http://news.xinhuanet.com/newmedia/2010-09/07/c_12527315.htm（查询时间：2010-09-08）