自动标注中文专利的引文信息

姜春涛
(南京大学计算机科学与技术系　南京 210023)
(江苏省专利信息服务中心　南京 210008)

摘要: 【目的】自动标注嵌入中文专利文本中的专利、标准、学术论文、其他专著 4 类引用信息。【方法】对于专利、标准和其他专著的引用，应用模式匹配的方法标注；对于学术论文的引用，应用由两阶段构成的机器学习方法标注，自动检测含有引用的句子，并从中自动提取 6 类文献特征信息。【结果】10 层交叉验证的结果表明: 专利引用标注的精确度和查全度均为 100%，标准引用标注的精确度和查全度分别达到 92%和 94%，而其他专著引用标注的精确度和查全度分别达到 80%和 71%；标注学术论文引用的精确度和查全度在阶段一分别为 95.7%和 96.0%，阶段二分别为 95.3%和 94.9%。【局限】模式匹配方法需要人工分析大量的专利文件，训练数据规模相对较小。【结论】运用模式匹配方法标注专利、标准引用的性能高于 92%；运用机器学习方法标注学术论文引用的平均性能达到 95%。

关键词: 专利引用文献提取　专利标注　模式匹配　条件随机场　信息提取
分类号: TP393

1　引　言

在专利信息挖掘的背景下，专利的引文信息作为衡量专利申请贡献程度的基础，对相关领域的研究具有重要的促进作用: 如现有技术搜索[1](Prior Art Search)，专利自动分类[2]，科学计量学[3](Scientometrics)等。然而，除了检索报告人工列举的引文信息，专利文件的正文主体(专利说明书)包含更多没有被检索报告列出的引文信息，如文献[1]指出一篇专利有可能包括上百个引用文献，而检索报告中所列出的引用文献却很少超过 10 篇。因此，从专利说明书中自动提取所引用的文献信息是十分必要的，而且极具价值[4]。

许多学者[4-5]都指出基于专利引文信息提取的研究工作还十分有限，需要更多深入的研究克服两方面的挑战: 所引用的专利文献书写形式的多样性及缺乏标准性；由于自然语言的歧义性和书写格式的多变性，提取所引用的非专利文献信息要比专利文献信息更加难以处理。尽管已有研究[1,5-9]从上述两方面针对英/日/德/法语专利文本进行有益的尝试，但是笔者没有发现对中文专利文本进行引文信息提取的研究。鉴于此，本文提出利用模式匹配和机器学习的方法，从中文专利说明书的“背景技术”和“具体实施方式”中自动标注所引用的专利和非专利文献信息。如果把所要标注的对象作为自动提取的内容，那么这种自动标注引文信息的过程，也可被视为自动提取引文信息的过程。

2　相关研究

文献[5-6]面向英文专利，提出使用模板或人工规则的方法自动提取专利和非专利文献信息，小规模的测试所提取非专利文献的精确度/查全度保持在 70%-75%。文献[7]使用正规表达式(Regular Expressions)建立大约 50 个引用专利文献的模式自动提取所引用的专利文献，然而进一步分析发现该方法会遗漏大约 40%的引用信息。文献[10]针对中文专利说明书摘要，使用规则和机器学习的方法，自动提取专利的特征、组成和用途信息。此外，针对中文专利信息自动提取的研究还包括面向本体的专利知识提取[11]，基于专利

通讯作者: 姜春涛，ORCID: 0000-0001-8332-7858，E-mail: spring_surge@126.com。

技术特征的聚类分析[12]，以及专利引文网络的可视化研究[13]，而从中文专利说明书中自动提取引文信息的研究还不多，这也被文献[14]的背景调查所证实。

条件随机场(Conditional Random Fields, CRF)[15]，作为一种序列标注(Sequence Labeling)算法的经典统计学模型，被成功应用于许多与自然语言处理相关的任务，如命名实体识别(Named Entity Recognition, NER)。文献[8]应用 CRF 模型从英文学术论文的参考文献中提取引用文献信息，其准确率达到 95.4%，比 HMM 模型高出 10%。文献[9]利用 CRF 模型从日本专利文本的背景技术中提取学术论文引用文献。文献[1]则训练 CRF 模型同时从英文专利的“背景技术”和“具体实施方式”中自动提取学术论文的信息。

3 研究方案

根据中文专利文本所引用文献的特点，采取具有针对性的方法进行引文信息的提取。笔者将嵌入在中文专利文本中的引用文献大致分为 4 类:专利、标准、学术论文、其他专著。对于专利、标准和其他专著的引用文献，因其引用形式有一定的规律可循，因此采用基于字符串模式匹配的方法自动提取，引用形式及正规表达式如表 1 所示；对于学术论文的引用文献，因其引用形式的多样性、易变性，则采用机器学习的方法处理。因应用模式匹配的方法相对简单，本文不作过多阐述，而侧重于应用机器学习的方法提取学术论文类引文信息，这也是本文的研究重点。

表 1 专利、标准、其他专著的引用形式及正规表达式

引用文献	引用形式	正规表达式
专利	“日本特开 2001-272593”	“日本特开 [0-9]{4}-[0-9]{6}”
专利	“美国专利 4,637,076”	“美国专利 [0-9],[0-9]{3},[0-9]{3}”
专利	“中国专利 CN02114474.5”	“中国专利 [A-Z]{2}[0-9]{8}.[0-9]”
标准	“EN10130-2006”	“EN[0-9]{5}-[0-9]{4}”
标准	“GB/T1539-1989”	“GB/T[0-9]{4}-[0-9]{4}”
标准	“G.657.A2”	“[A-Z].[0-9]{3}.[A-Z0-9]{2}”
其他专著	“SDP: Session Description Protocol”	“[“].*[”]”
其他专著	“《网络通信技术》”	“[《].*[》]”

3.1 引用文献特性

专利说明书中，对相关专利的引用通常包括两部分信息:

(1) 签发机构名称，以国家代码(如 DE、US)或区域代码(如 EP、WO)表示。

(2) 因签发机构而异的专利编码。签发机构可由文字表示也能以编码表示，如“US2004/0208331”，“日本专利 JP-A-10-224951”。同样地，签发机构即使在同一篇专利文本中，也可能由不同的名称来引用，如“日本特开平 11-61327 号公报”、“实公昭 56-23294 号公报”。

对标准的引用形式与对专利的引用大同小异，专利撰写者通常同时使用文字和编码，如“《城镇污水处理厂污染物排放标准》(GB18918-2002)”。对学术论文的引用，归纳为专著、期刊、学术会议论文、硕博论文、技术报告 5 种类型。对其他专著的引用，概括为对技术手册、草案等非学术类论文文献的引用。

3.2 模式匹配方法

基于字符串模式匹配方法的思路是利用自然语言的语义和句法构建符合所要提取的引用文献的模式库，不同形式的引用文献由不同的模式描述。对专利、标准及其他专著而言，笔者通过使用符合 IEEE POSIX 句法标准的正规表达式，表 1 列出了如何构建不同类型的引用模式。

3.3 机器学习方法

表 2 CRF 模型所使用的特征及特征组合

特征	组成	特征组合	组成
F1	目标词	F17	F4/F1
F2	目标词的词性	F18	F1/F10
F3	目标词是否线索词	F19	F8/F5
F4	目标词前第一个词	F20	F5/F2
F5	目标词前第一个词的词性	F21	F2/F11
F6	目标词前第一个词是否线索词	F22	F11/F14
F7	目标词前第二个词	F23	F8/F5/F2
F8	目标词前第二个词的词性	F24	F5/F2/F11
F9	目标词前第二个词是否线索词	F25	F2/F11/F14
F10	目标词后第一个词	F26	F9/F6
F11	目标词后第一个词的词性	F27	F6/F3
F12	目标词后第一个词是否线索词	F28	F3/F12
F13	目标词后第二个词	F29	F12/F15
F14	目标词后第二个词的词性	F30	F9/F6/F3
F15	目标词后第二个词是否线索词	F31	F6/F3/F12
F16	目标词前第一个词的输出标签	F32	F3/F12/F15
		F33	F2/F3

受文献[9]所提出方法思路的启发,笔者将自动提取学术论文类引用文献的任务分为两个阶段:利用自动分类方法从专利文本中自动提取包含有引用文献信息的句子; 利用 CRF 模型从包含引用文献信息的句子中自动提取相应的引用文献特征信息, CRF 模型所使用的特征及特征组合如表 2 所示。

(1) 提取含有引用文献的句子

含有引用文献的句子在其所引用之处或其上下文, 通常会出现一些线索词, 如‘发表’、‘公开’、‘第 x 卷 x 号’、‘Proceedings’、‘pages’、‘Transactions’等。假设专利文本被分解为含有两个类别的句子集合: 即含有引用文献信息的句子和完全不含有引用文献信息的句子, 那么自动提取含有引用文献的句子的任务即可转化为自动二分类的任务。

笔者从表 3 的 A-H 的各个 IPC 类别的专利文件中随机选取 50-200 篇, 共计 2 192 篇专利文件, 并从中提取“背景技术”和“具体实施方式”两部分的全文本, 人工选取含有引用文献的句子前后出现的线索词, 共计 174 个, 并把这些线索词作为描述专利文本特征向量的特征, 其值由是否出现在句子中所决定(即出现为 1, 反之为 0)。这样, 在由这些特征所构成的专利文本特征向量上应用自动分类的算法(如 SVM)[16]训练分类模型, 含有引用文献的句子则可以被自动标记。

表 3　中文专利实验数据统计分布

IPC	专利	件数	IPC	专利	件数
A41	服装	855	D21	造纸/纤维素	733
A42	帽类制品	101	E02	水利工程/基础/疏浚	1 827
A43	鞋类	604	E03	给水/排水	665
A45	手携/旅行物品	869	E05	锁/钥匙/门窗零件等	1 150
B08	清洁	813	F21	照明	3 920
B22	铸造/粉末冶金	2 394	F24	供热/炉灶/通风	3 186
B24	磨削/抛光	1 433	F26	干燥	398
B30	压力机	321	F27	炉/烘烤炉/蒸馏炉	485
C02	水/废水/污水等	3 636	G02	光学	5 443
C03	玻璃/矿棉/渣棉	1 417	G08	信号装置	1 501
C21	铁的冶金	1 860	H02	发电/变电/配点	2 405
D01	天然/人造的线/纤维等	1 114	H03	基本电子电路	317
D03	织造	498	H04	电通信技术	6 104
D05	缝纫/绣花/簇绒	295	总计		44 344

(2) 提取引用文献特征信息

笔者应用 CRF 从第一阶段的结果中自动提取 6 种引用文献特征信息: 题目、作者、文献源、日期、卷标、页数。与其他标注算法模型(HMM,SVM)相比, CRF 模型可以使用丰富的特征信息, 任意数量的上下文信息, 并确保其计算结果收敛于全局最优。此外, 相关文献[8-9]分别从其实验中证实 CRF 模型的标注结果的平均性能比 HMM 模型高 13.9%, 比 SVM 模型高 4.6%。因此本文使用 CRF 模型, 其思路如下: 如果把每个含有引用文献的句子看作一个字符序列, 那么从句中自动提取引用文献特征的过程即可转换为使用 CRF 对字符序列进行自动标注的过程, CRF 所标注的内容即为所要提取的引用文献特征。笔者定义下列标签用以标记所要提取的引用文献特征: A-作者; T-文献题目; S-文献源(包括期刊或学术会议的名称、出版社等); D-出版日期; P-页数; VN-卷标。通过 IOB2 编码[17], 可得到如下标签集:

{B-A,I-A,B-T,I-T,B-S,I-S,B-D,I-D,B-P,I-P,B-VN, I-VN,O}

其中, B 表示标记的开始, I 表示标记之内, O 表示标记之外。笔者使用这 13 个标签人工标记经过分词、词性标注预处理的含有引用文献的句子作为训练数据, 从而训练 CRF 模型, 用以标注引用文献特征。本文选取的特征组合包括 F1-F33 共 33 个特征, 其中两个或两个以上的特征由‘/’连接(见表 2)。

4　实验结果

本文实验环境为 CPU: Intel Core i5-3470 3.2GHz, 内存: 4GB, 操作系统: Fedora Linux 20, 编程语言: Oracle Java 8。

4.1　专利数据集

从国家区域中心数据库下载 2010 年-2014 年的 IPC 类 A-H 的专利授权文件共计 44 344 件, 其统计分布如表 3 所示, 并用 Java 编制程序自动提取专利说明书的“背景技术”和“具体实施方式”的全文本, 进行相应的预处理, 以此作为实验数据集。

4.2　自动提取专利、标准及其他专著

(1) 测试数据

从表 3 列出的 44 344 个专利文件的文本中, 人工提取对专利、标准及其他专著引用的字符串模式, 用以建立引用模式库。从 IPC 类别 A-H 随机选取 50-100

个专利文件，共组成 2 828 个专利文件，提取相应的“具体实施方式”文本，并人工标记其中包含的专利引用(234 个)、标准引用(255 个)、其他专著引用(28 个)作为测试的标准。对于测试结果的性能，笔者使用经典评测方法 Precision(精确度)和 Recall(查全度)，其定义如下:

$$\text{Precision} = \frac{\text{正确提取的引用文献的数目}}{\text{使用模式匹配方法提取的引用文献的数目}}$$

$$\text{Recall} = \frac{\text{正确提取的引用文献的数目}}{\text{人工标记的引用文献的数目}}$$

(2) 测试结果和讨论

通过使用 Java Regular Expressions API①，笔者构建针对专利、标准、其他专著引用的模式库，其中包括 17 条专利引用模式，12 条标准引用模式，6 条其他专著引用模式。这样，通过字符串模式匹配而得到的结果如表 4 所示:

表 4　模式匹配方法的提取结果

引用文献	Precision	Recall
专利	100%	100%
标准	92%(240/261)	94%(240/255)
其他专著	80%(20/25)	71%(20/28)

由表 4 可看出，对专利引用的提取精确度和查全度最高，标准引用次之，其他专著引用最低。对于专利和标准而言，它们引用模式的规则性、重复性很强，有助于利用正规表达式构建匹配模式库，在构建模式匹配库时所使用的数据规模很大，分布很广的前提下，经过细致筛选、匹配可以取得很好的性能。然而，对于其他专著来说，一部分无规则可循的引用，其形式随意性较大，缩写被频繁使用，撰写者通常只列出题目、作者、公司名称的一种。例如，“参见 USP 24, 2000 版，第 19-20 页和第 856 页(1999)”，此引用是本文模式匹配方法没有检测到的。实际应用中，在缺少大量数据的情况下，模式库无法包含所有的匹配模式，因而查全度很难获得明显的提升。鉴于此，笔者认为使用机器学习的方法处理模式匹配方法所遗漏的引用文献，是一种更好的补充选择。

4.3　自动检测含有学术论文引用的句子

(1) 测试数据

笔者从表 3 中 A-H 各个类别的专利文件中随机选取 50-200 篇，共计 2 192 篇，并从“背景技术”和“具体实施方式”的全文本中选取 19 746 个句子进行人工标注: 17 458 个句子作为训练数据(其中 1 764 个句子被标记为含有学术论文引用)，2 288 个句子作为测试数据(其中 300 个句子被标记为含有学术论文引用)。同样，使用 Precision 和 Recall 作为评测的标准，其定义如下:

$$\text{Precision} = \frac{\text{正确检测的句子的数目}}{\text{使用分类方法所检测的句子的数目}}$$

$$\text{Recall} = \frac{\text{正确检测的句子的数目}}{\text{人工标记含有引文信息的句子的数目}}$$

(2) 测试结果和讨论

经过测试多个自动分类算法模型(Naive Bayesian、Decision Tree 和 SVM)，发现 SVM 的分类结果最佳。因此笔者采用开源机器学习软件 Weka[18] 中的 SVM 算法 LibLinear[19] 在含有 17 458 个句子的训练数据上构建 SVM 检测器模型，表示为“SVM 检测器”，并使用含有 2 288 个句子的测试数据集进行测试，结果如表 5 所示。同时列出文献[9]所采用的方法结果，表示为“TinySVM 检测器”。

表 5　使用 SVM 检测器的结果

比较项	SVM 检测器	TinySVM 检测器
特征	174	36
Precision	95.7%(288/301)	91.6%(252/275)
Recall	96.0%(288/300)	86.9%(252/290)

需要指出的是: TinySVM 检测器应用于日本专利文本，而笔者的 SVM 检测器针对于中文专利文本。此外，SVM 检测器和 TinySVM 检测器在检测范围、模型的构建以及算法的选择上有显著的不同:

①前者能同时处理中文专利的“背景技术”和“具体实施方式”两部分的中、英文本，而后者只处理日本专利的“背景技术”的日、英文本;

②前者使用的特征数为 174，远大于后者所选取的 36;

③在先期实验中，前者采用的线性 SVM 算法，其平均精确度和平均查全度要比后者所采用的基于多项式核函数(Polynomial Kernel)的 SVM 算法分别高 23.5%和 12.1%。

4.4　自动提取引用特征信息

(1) 测试数据

笔者从 4.3 节人工标注的测试数据中，选取 450

①https://docs.oracle.com/javase/8/docs/api/java/util/regex/package-summary.html.

个含有学术论文引用的句子，首先使用 Stanford Word Segmenter①对每个句子进行分词，然后使用 Stanford POS Tagger②对分词结果进行词性标注，根据 3.3 节的方法对每个句子进行人工标注，作为测试数据。同样，利用 Precision 和 Recall 衡量 CRF 训练模型的性能，其定义与 4.3 节的定义类似，故不再赘述。

(2) 测试结果和讨论

笔者应用开源 CRF 软件包(CRFsuite)[20]，对 450 个句子进行 Ten-Fold Cross Validation(10 层交叉验证)的测试，其标注结果与文献[9]采用 CRF++的方法进行比较，如表 6 所示。其中，–表示对于卷标这一特征，CRF++方法无法提取。

表 6　使用 CRF 模型标注的结果

文献特征	本文的方法		文献[9]的方法	
	Precision	Recall	Precision	Recall
题目	99.4%	99.5%	英: 74.6% 日: 84.8%	英: 90.3% 日: 88.1%
作者	98.2%	96.8%	英: 87.2% 日: 88.5%	英: 85.7% 日: 76.5%
文献源	97.3%	95.9%	英: 80.5% 日: 83.4%	英: 79.9% 日: 73.6%
日期	90.4%	92.4%	93.2%	92.1%
卷标	93.1%	92.8%	–	–
页数	93.6%	92.3%	97.3%	97.3%
平均	95.3%	94.9%	86.2%	85.4%

由表 6 可知，使用本文的方法标注 6 个文献特征信息的平均精确度和平均查全度分别达到 95.3%和 94.9%，其中对于题目的识别率最高，作者和文献源的识别率相接近，稍微低于题目的识别率；卷标和页数的识别率高于日期的识别率，这是因为前两者的上下文中有利于识别的线索词的出现，而后者仅仅是数字，且在引用文本中出现的位置不固定，很难与其他相类似信息区分。与 CRF++方法[9]相比较，本文的方法在训练 CRF 模型所使用的特征、CRF 具体的训练算法及 CRF 模型所能标注的文献特征信息有显著不同:

①前者所使用的特征为目标词前后两个词的特征以及目标词前两个词的输出标签，而后者所使用的特征则由表 2 所示;

②前者采用经典 L-BFGS[21] 训练算法，而后者则采用 Passive Aggressive[22] 训练算法;

③前者是把英、日文引用文献信息分开标注，并且没有标注卷标特征，而后者则混合标注中、英文引用文献特征信息，并且单独标注卷标特征。

5　结　语

根据嵌入中文专利文本的引文信息的特性，本文提出利用模式匹配的方法提取专利、标准及其他专著的引用，利用机器学习的方法提取学术论文的引用。由 27 个技术领域的中文专利构成的数据集测试结果可知: 提取专利和标准的精确度和查全度已达到 92%以上，而提取其他专著的查全度则只有 71%，这与部分其他专著的引用形式多样化且无规则性有关；检测含有引文信息的精确度和查全度比采用类似方法分别提高 4.1%和 9.1%，而提取 6 种文献特征的平均精确度和查全度则要分别高于所比较的方法 9.1%和 9.5%。经过实验结果的分析，笔者考虑在今后的研究中结合模式匹配和机器学习的方法提取非专利引文信息。此外，将扩大测试数据的领域覆盖度及人工标注集的规模，从而为后续研究提供有力的支持。

参考文献：

[1] Lopez P. Automatic Extraction and Resolution of Bibliographical References in Patent Documents [A]. // Advances in Multidisciplinary Retrieval [M]. Springer Berlin Heidelberg, 2010: 120-135.

[2] Lai K K, Wu S J. Using the Patent Co-citation Approach to Establish a New Patent Classification System [J]. Information Processing and Management, 2005, 41(2): 313-330.

[3] Mayer M. Does Science Push Technology? Patents Citing Scientific Literature [J]. Research Policy, 2000, 29(3): 409-434.

[4] Adams S. The Text, the Full Text and Nothing but the Text: Part 1-Standards for Creating Textual Information in Patent Documents and General Search Implications [J]. World Patent Information, 2010, 32(1): 22-29.

[5] Lawson M, Kemp N, Lynch M F, et al. Automatic Extraction of Citations from the Text of English-language Patents - An Example of Template Mining [J]. Journal of Information

①http://nlp.stanford.edu/software/segmenter.shtml.

②http://nlp.stanford.edu/software/tagger.shtml.

Science, 1996, 22(6): 423-436.

[6] Agatonovic M, Aswani N, Bontcheva K, et al. Large-scale Parallel Automatic Patent Annotation [C]. In: Proceedings of the 1st ACM Workshop on Patent Information Retrieval. ACM, 2008.

[7] Lopez P, Romary L. Multiple Retrieval Models and Regression Models for Prior Art Search [C]. In:Proceedings of the 2009 Cross-Language Evaluation Forum Workshop. Springer, 2009.

[8] Peng F, McCallum A. Accurate Information Extraction from Research Papers Using Conditional Random Fields [C]. In: Proceedings of the 2002 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL). 2004: 329-336.

[9] Nanba H, Anzen N, Okumura M. Automatic Extraction of Citation Information in Japanese Patent Applications [J]. International Journal on Digital Library, 2008, 9(2): 151-161.

[10] Feng G, Chen X, Peng Z. A Rules and Statistical Learning Based Method for Chinese Patent Information Extraction [C]. In: Proceedings of the 8th Web Information Systems and Applications Conference. IEEE, 2011:114-118.

[11] 姜彩红，乔晓东，朱礼军．基于本体的专利摘要知识抽取[J]．现代图书情报技术，2009(2): 23-28. (Jiang Caihong, Qiao Xiaodong, Zhu Lijun. Ontology-based Patent Abstracts' Knowledge Extraction [J]. New Technology of Library and Information Service, 2009(2): 23-28.)

[12] 王曰芬，徐丹丹，李飞．专利信息内容挖掘及其试验研究[J]．现代图书情报技术, 2008(12): 59-65. (Wang Yuefen, Xu Dandan, Li Fei. Experimental Study of Patent Information Content Mining [J]. New Technology of Library and Information Service, 2008(12): 59-65.)

[13] 于霜．基于专利引文网络的空间关系可视化研究[D]．大连：大连理工大学，2010. (Yu Shuang. Analysis on Visualization Among Spatial Relationship Based on Patent Citation Network [D]. Dalian: Dalian University of Technology, 2010.)

[14] 薄怀霞．基于构建专利引文数据库的专利文献分析研究[D]．曲阜：曲阜师范大学，2014. (Bo Huaixia. Patent Literature Analysis Study Based on Building Patent Citation Databases [D]. Qufu: Qufu Normal University, 2014.)

[15] Lafferty J D, Mccallum A, Pereira F C N. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data [C]. In: Proceedings of the 18th International Conference on Machine Learning. 2001: 282-289.

[16] Vapnik V N. The Nature of Statistical Learning Theory [M]. The 2nd Edition. Springer, 1999.

[17] Cho H C, Okazak N, Miwa M, et al. Named Entity Recognition with Multiple Segment Representations [J]. Information Processing and Management, 2013, 49(4): 954-965.

[18] Hall M, Frank E, Holmes G, et al. The WEKA Data Mining Software: An Update [J]. SIGKDD Explorations, 2009, 11(1): 10-18.

[19] Fan R E, Chang K W, Hsieh C J, et al. LibLinear: A Library for Large Linear Classification [J]. Journal of Machine Learning Research, 2008, 9(12): 1871-1874.

[20] Okazaki N. CRFsuite: A Fast Implementation of Conditional Random Fields [CP/OL]. [2015-03-24]. http://www.chokkan.org/software/crfsuite/.

[21] Nocedal J. Updating Quasi-Newton Matrices with Limited Storage [J]. Mathematics of Computation, 1980, 35(151): 773-782.

[22] Crammer K, Dekel O, Keshet J, et al. Online Passive Aggressive Algorithms [J]. Journal of Machine Learning Research, 2006, 7(3): 551-585.

收稿日期: 2015-04-14
收修改稿日期: 2015-06-11

Automatic Annotation of Bibliographical References in Chinese Patent Documents

Jiang Chuntao
(Department of Computer Science and Technology, Nanjing University, Nanjing 210023, China)
(Patent Information and Service Center of Jiangsu Province, Nanjing 210008, China)

Abstract: [**Objective**] This paper aims to automatically annotate four types of bibliographical references in Chinese patent documents, such as patents, standards, papers, and other monographs public documents. [**Methods**] Use a pattern matching approach to annotate the references of patents, standards, and public documents, and use a two-phase machine learning approach to annotate the paper references, firstly, automatically detecte the sentences that contain citation information, then extracte 6 categories of bibliographic features from the results. [**Results**] The results of ten-fold cross validation show that the accuracy for annotating patents is 100%, and the precision and recall for annotating standards is 92% and 94% respectively, while the precision and recall for annotating public documents is 80% and 71% respectively. For annotating paper references, the precision and recall in phase one is 95.7% and 96.0% and in phase two is 95.3% and 94.9% respectively. [**Limitations**] The pattern matching approach requires analyzing a lot of patent documents manually, and the size of the training model used by the proposed machine learning approach is relatively small. [**Conclusions**] The performance of annotating patents and standards using a pattern matching approach achieves over 92%, and the performance of annotating papers using a machine learning approach achieves 95%.

Keywords: Patent citation extraction　Patent annotation　Pattern matching　Conditional Random Fields
Information extraction

康奈尔大学图书馆与 ProQuest 合作改善图书采购效率

ProQuest 正与康奈尔大学图书馆展开合作，试图开发动态的图书选择系统，将来自多源头的图书元数据聚集到一个单一的简化界面，从而提高图书采购工作效率。简化界面是康奈尔大学图书馆新协议的一部分，这一新协议规定康奈尔大学图书馆将使用 ProQuest 作为其英语语种的印刷图书的主要来源，这是康奈尔大学图书馆与 Coutts 信息服务商长期以来的需求驱动的图书采购合作计划的扩展。

“随着图书采购选择范围的不断扩大，我们需要能够帮助图书馆员无需从各种不同系统进行筛选就能做出明智决策的辅助工具。”康奈尔大学图书馆采购和电子资源许可服务主管 Jesse Koennecke 说：“很高兴能够与 Coutts 和 ProQuest 建立长期合作关系，来帮助改善图书采购的体验。”

“ProQuest 和康奈尔大学图书馆都认为图书馆和供应商保持良好的合作关系对于促进创新是非常重要的。” ProQuest 副总裁 Bob Nardini 表示：“我们很高兴能与一批高度敬业的图书馆员合作，共同打造一个优秀的图书选择系统，为全世界的图书馆谋福利。”

在 OASIS(Online Acquisitions and Selection Information System)中，新的界面能够将来自多个源头的图书元数据聚集成一个单一的选择和采购流程。OASIS 是一个基于网络的系统，能够搜索、选择和订购印刷本及电子版图书，其强大的检索工具能检索市场上绝大多数的数据库并快速精准地定位到对应的条目。

Coutts 信息服务商于 2015 年 4 月加盟 ProQuest，将其在馆藏建设、图书编目方面的专长，以及 MyiLibrary 和 OASIS 平台带入了 ProQuest。Coutts 是 ProQuest 图书部分不可或缺的一员，拥有一个快速发展的电子书内容发现、访问和管理技术框架。

(编译自: http://www.proquest.com/about/news/2015/Cornell-University-Library-and-ProQuest-Team-Up-for-Faster-More-Efficient-Book-Selection-and-Acquisition.html)

(本刊讯)