科技文献元数据自动抽取研究述评①

龚立群, 马宝英, 常晓荣

(昌吉学院 计算机工程系, 昌吉 831100)

摘 要: 首先从元数据的属性和元数据的粒度两个角度对科技文献元数据进行了分析, 在此基础上, 从科技文献元数据自动抽取的理论研究和应用实践研究两个方面对国内外科技文献元数据自动抽取研究成果进行分析和综合, 最后指出了现有研究的特点和存在的不足.

关键词: 科技文献; 元数据自动抽取; 基于规则的抽取; 基于模板的抽取; 基于机器学习的抽取

Literature Review on Automatic Metadata Extraction of Scientific Paper

GONG Li-Qun, MA Bao-Ying, CHANG Xiao-Rong

(Computer Engineering, Changji College, Changji 831100, China)

Abstract: From the perspectives of metadata attributes and metadata granularity, the metadata of scientific paper is analyzed. On this basis, the research on metadata extraction of scientific paper in domestic and international are analyzed and synthesized from two aspects of the theoretical research and application in practice. Finally, the features and shortcomings of the current research are pointed out.

Key words: scientific paper; automatic metadata extraction; rule-based extraction; template-based extraction; machine-learning extraction

在传统的图书馆中, 文献的元数据信息(如标题、作者、参考文献等)往往由文献的生产者(作者)或加工者(图书馆员)手工抽取或录入的. 但随着目前网络上的科技文献数量激增, 单靠人工抽取或录入这些元数据已不太可能, 另外, 大量的遗留纸质文档中的信息在转化为数字文档的过程中, 也需要能够自动抽取这些文档中的元数据.

元数据自动抽取是信息抽取(Information Extraction, IE)的研究内容之一, 科技文献元数据的自动抽取能够充分利用科技文献本身所具有的内在结构信息来实现信息抽取, 可以看作是面向领域的信息抽取.

对数字图书馆中大量的异构的科技文献实现其元数据的自动抽取是一项具有挑战性的工作. 在这里, 科技文献的异构性主要体现在如下几个方面: 科技文献的版面布局的不同; 科技文献可能包含不同的元数据元素, 如某一文献可能包括标题、作者、摘要、日期等元数据, 而另一文献则可能包含标题、作者、出版者等元数据; 不同文献中的元数据的出现顺序可能不同.

近年来, 国内外学术界对科技文献元数据的自动抽取展开了相应的研究(如基于机器学习的元数据自动抽取研究、基于规则的元数据自动抽取研究), 业界也设计和开发了一些科技文献元数据自动抽取工具(如 Metadata Miner Catalogue Pro、MetadataExtractor 等). 本文在综述国内外科技文献元数据自动抽取理论研究和应用实践研究的基础上, 探讨科技文献元数据自动抽取的研究进展, 并指出现有研究的特点和存在的不足.

1 科技文献元数据概述

本文从元数据的属性和元数据的粒度两个角度对科技文献元数据进行分类.

1.1 从元数据的属性来看

从元数据的属性来看, 元数据一般分为管理性元

① 基金项目:教育部人文社会科学研究规划基金(09XJA870003)

收稿时间:2012-09-01;收到修改稿时间:2012-09-27

数据、描述性元数据技术性元数据和使用性元数据. 目前的元数据自动抽取工具对于管理性元数据、技术性元数据的抽取效果较好, 而对于描述性元数据的抽取效果则较差. 研究者从理论和实验角度所进行的研究大多针对的是描述性元数据的抽取.

1.2 从元数据的粒度来看

科技文献中的元数据按照其粒度大小可以分为单一对象元数据和复杂对象元数据. 单一对象元数据是指基本身不再进行进一步分割的元数据对象, 如对科技文献头部的标题、作者等元数据; 复杂对象元数据是指本身还需要进一步分割的元数据对象, 如科技文献尾部参考文献、科技文献中的图、表、公式等元数据对象. 复杂对象元数据的自动抽取要比单一对象元数据的自动抽取困难. 早期的一些研究主要是研究如何自动抽取科技文献中的单一对象元数据.

2 国内外相关的研究

国内外关于科技文献元数据抽取的研究分为理论研究和应用实践(元数据抽取工具)研究.

2.1 理论研究

国内外对于科技文献元数据的抽取的理论研究按照所采用的方法可以分为基于规则的元数据抽取研究、基于模板的元数据抽取研究和基于机器学习的元数据抽取研究.

2.1.1 基于规则的科技文献元数据抽取

基于规则的元数据抽取使用一组由领域专家事先定义好的规则来抽取元数据. 在基于规则的方法中, 规则是关于每个字段的一般知识, 每个规则都与某个值相关, 来显示其可能性. 这些规则主要是基于目标文档的视觉线索来定义. 例如科技文献的标题元数据一般字体最大, 位置最靠上, 利用这些悔改特征和位置特征来编写规则, 从而定位和抽取标题元数据.

Wei Wei 等人[1]提出了一种基于规则的双层标注方法来抽取相引文元数据, 他们所采用的方法在分析引文元数据的外观和标点符号的基础上, 在两个事先定义的解析层执行格式标注和语义标注. Besagni 等[2]提出的方法也可以认为是基于规则的方法, 他们通过词性标注的方法标注引文元数据的片段. 李朝光等[3,4]利用正则表达式规则对论文元数据信息进行信息抽取, 这种方法充分利用论文特有的结构, 在不采用语法分析等复杂的自然语言处理手段的情况下取得了很好的效果. 陈俊林等[5]通过把 PDF 文件用工具 PDF2HTML 工具转换成中间文档, 再总结出标题、作者名、作者地址、E-mail 共四类论文元数据特征, 最后利用 XSLT 作为抽取规则制定语言进行抽取, 但是在其研究中元数据类型不够丰富, 特征总结不够全面, 还有待于进一步研究.

由于基于规则的方法实现简单且准确率较高, 目前大部元数据抽取系统都是基于规则的元数据抽取系统. 例如最为典型的 CiteSeer 系统利用启发式规则自动抽取 PDF 或 Postscript 格式的学术论文的元数据, 并且提供某一具体文献的“引用”和“被引”情况和相关文献, 成为当前流行的学术搜索引擎之一.

基于规则的方法缺点是需要事先由领域专家设计一系列的抽取规则, 并要时时对这些规则进行维护, 另外抽取规则的适应性较差, 而不同的科技文献的格式往往不同, 甚至当有较多的规则存在时, 还需要解决规则间的不一致性和冲突, 文献的特征数量越多, 所需要制定的规则数量就越多, 这使得基于规则的系统难以处理特征数量较多的文献资料. 基于规则的元数据抽取的另一个缺点是一旦规则被定义后, 就固定了下来, 所以当错误发生时, 对系统进行调整就比较困难.

2.1.2 基于模板的元数据抽取

基于模板的元数据抽取方法使用由各种风格的引文模板所组成的模板库来抽取元数据. 这种方法一般先建立模板数据库, 然后查找和匹配模板, 完成待匹配元数据的抽取. 例如如果某条参考文献的样式是(作者), “(标题名)”, (期刊名称), (时间), (页码), 则可以用如下的模板来表示这种样式: ‘_AUTHOR_, “_TITLE_”,_JOURNAL_,_DATE_,pp._PAGES_.’.

Ding 等人[6]利用自然语言文本中的基于模式识别和模式匹配的模板挖掘方法从数字文档中抽取不同类型的信息. Min-Yuh Day 等[7]采用一种层次化的知识表示框架来抽取科技文献尾部的参考文献元数据, 他们认为一般的基于规则的元数据抽取方法是“扁平的”, 而他们所提出的基于模板的方法则是层次化的, 这种层次化的知识表示框架(INFOMAP)提供了一个集成的层次化模板编辑环境及一个灵活的模板匹配引擎, 实验表明, 其抽取效果要优于一般的基于模板元数据抽取方法. Eli Cortez[8]提出了一种无监督的引文元数据抽取方法(FLUX-CIM)来从训练样本集中自动产生模板. Kai-Hsiang Yang 等人[9]设计了一个基于序列队列技术的引文元数据解析工具 BIBPRO. 郭志鑫等[10]提出了一种

从科技文献尾部提取参考文献元数据信息的方法, 这种方法采用模板匹配方式(系统中共有预先设计好的模板1247 个, 并且可以动态地增加), 从文档中提取作者、标题、出版时间等信息, 进一步地利用语义网中的本体理论, 使用 OWL 本体描述语言对提取出的元数据信息进行格式化, 实现信息的语义表示. 高良才等[11]在总结现在引文元数据抽取方法的基础上, 针对引文的排版惯例——引文在文档内部风格一致, 提出了一种新的基于模板的引文元数据抽取方法, 他们的实验结果表明此方法在引文元数据发现、分割和标注方面取得了较好的效果.

基于模板的元数据抽取方法需要解决的问题主要有两个, 一个是模板库的构建, 另一个是模板的匹配过程. 基于模板的元数据抽取方法实现简单, 但其元数据抽取结果严重依赖于数字文档的风格和版式.

2.1.3 基于机器学习的科技文献元数据抽取

基于机器学习的元数据抽取方法利用机器学习技术来抽取元数据, 例如 HMM(Hidden Markov Model)模型、SVM(Support Vector Machine)模型和 CRF (Conditional Random Fields)模型等.

在基于机器学习的方法中, 最先应用于科技文献元数据抽取的是隐马尔可夫 HMM 模型. 基于 HMM 模型的元数据抽取把文档看作是由一些隐藏状态产生的词组序列, 从中找出最可能的状态序列. Seymore 等[12]利用 HMM 模型来从计算机领域文献的头部抽取重要的字段. Junfei Geng[13]利用基于 HMM 模型的解析工具(AutoBib)解析从网上抽取的计算机科学领域的原始记录来形成引文元数据. HMM 模型不需要大规模的词典集和规则集, 具有学习能力且可适应性好, 但对于非独立特征的建模存在一定的困难.

后来又有研究者采用 SVM 模型来抽取元数据信息. SVM 方法将元数据抽取看作是使用相应的元标记(metatag)对文本进行标注, 每个元标记对应于一个类别, 这样元数据抽取问题就可以采用分类的方法来解决. H.Han 等[14]描述了一种基于 SVM 分类的方法从研究文献的头部抽取元数据, 他们的实验结果表明基于 SVM 的元数据抽取算法要优于 Seymore 等的基于 HMM 的算法. 欧阳辉和禄乐滨[15]通过分析多分类 SVM 的特点, 建立了基于平衡二叉树的支持向量机模型 BBT-SVM, 并在训练过程中调整相关参数, 得到目标支持向量机, 针对 PDF 文档的特点, 采用 PDFbox 开源库对 PDF 文档进行解析, 去除 PDF 文档的文件头等额外文档描述信息, 得到目标信息, 最后利用 libsvm 开源库对 PDF 目标信息进行元数据抽取, 实验结果表明各类元数据的查全率都在86%以上, 查准率都在92%以上, 效果较好. 基于 SVM 的元数据抽取的优点是与 HMM 方法相比, 具有较好的抽取性能. SVM 元数据抽取的训练样本小, 学习速度快, 易于扩展, 缺点是缺失了状态转移和观察序列之间的紧密关系, 较难选择正确的特征, 对大量训练集数据标注也较困难, 训练过程比较耗时.

也有研究者将 SVM 和 HMM 算法结合起来进行元数据抽取. Takasu 等人[16,17]综合使用 SVM 算法和 HMM 算法来处理参考文献中的词频(如果参考文献中的字段包括数字型数据, 如 2004, 则需要使用词频)和语法信息. 张铭等人[18]提出了 SVM+BiHMM 的元数据抽取模型, 首先根据训练集分别建立独立的 SVM 模型和 BihMM 模型, 采用 Sigmoid 双弯曲函数把 SVM 分类结果拟合为 BiHMM 模型的单词发射概率, 再采用 SVM+BiHMM 复合模型进行元数据抽取, 改善了单独利用 HMM 和 SVM 方法进行论文元数据抽取的效果.

条件随机域方法允许使用任意和独立特征和全序列上的交叉引用特征, 兼有 HMM 和 SVM 的优点. Peng 和 McCallum 等[10]采用 CRF 来抽取科技文献的头部元数据(包括标题、作者、机构等 15 个元数据)和科技文献尾部的参考文献元数据(包括作者、标题、编者等 13 个字段信息), 他们的实验结果表明基于 CRF 的元数据抽取算法要优于基于 HMM 和基于 SVM 的元数据抽取算法. Yu 等[20]在中文科技论文数据集上测试了利用 CRF 方法抽取论文头部和引文元数据的方法, 同样取得了良好的效果. 现有的研究表明, 在 HMM、SVM 和 CRF 三种方法中, 在同一测试数据集上对于论文头部元数据的抽取, CRF 的抽取的总体精确度最高, 达到 98.3%(HMM 是 93.1%, SVM 是 92.9%)CRF 的缺点是需要较长的训练时间.

基于机器学习元数据抽取的方法鲁棒性和可适应性较好, 但事先需要人工标注的语料集的训练, 而对大量样本进行标注和训练比较费时、费力.

2.2 应用实践研究

根据 ERPANET 的 packaged object ingest project[21]项目显示目前只有很少的几个元数据自动抽取工具, 而且大部分只是用来抽取技术性元数据(如 DROID[22]和新西兰国家图书馆的 metadata Extraction Tool[23]). 尽管目前也有一些项目在提供元数据抽取工具(如 Waterloo 大

学和 DC Initiative 的 MetadataExtractor[24], Catholic 大学的 Automatic Metadata Generation[25]和由 Soft Experience 开发的商业软件 Metadata Miner Catalogue Pro[26])来抽取有限的几个描述性元数据(如标题、作者和关键词), 但这些工具依赖于结构化的文档, 因而其精确性和有效性是有限的. 同时, 目前也没有自动化的元数据抽取工具来抽取高层的语义元数据(如内容摘要).

CiteSeer[27]是索引电子学术文献的自动引文索引系统的一个例子, 它使用基于规则方法抽取各种形式的参考文献, CiteSeer 抽取标题和作者的精确度大约是 80%, 抽取文献页码的精确度大约是 40%.

3 现有研究的分析和评述

通过分析发现现有的研究存在着以下特点及不足.

3.1 理论或实验研究与实践工具开发脱节

目前在元数据抽取领域大多所进行的是一些实验性的研究, 尽管也有一些元数据抽取工具在实际操作环境中抽取元数据, 但这些元数据抽取工具在开发时并没有充分利用实验研究结果, 因此, 在元数据抽取领域实验研究和应用开发之间存在着脱节现象. 如果今后将实验研究与应用开发活动充分地整合起来, 将会极大地促进元数据自动抽取技术的发展.

3.2 复杂对象与单一对象的抽取

早期的元数据抽取研究主要研究单一对象的抽取, 虽然也抽取文献尾部的引文信息, 但把一篇文献的所有引文作为一个整体进行处理. 相比单一对象的抽取, 复杂对象的自动抽取要更难一些. 后来的一些研究开始研究复杂对象的抽取, 如抽取文献中的表格、文献尾部的引文信息等.

3.3 元数据抽取方法的综合应用

现有的科技文献元数据抽取大多仅使用一种方法抽取元数据, 即或者仅使用基于规则的方法, 或者仅使用基于机器学习的方法, 或者仅使用基于模板的方法. 而且在基于机器学习的抽取研究中, 目前的研究者大多是采用单一的学习模型(例如 HMM、SVM 等)来进行元数据抽取.

就基于规则的抽取方法来看, 存在着一定的问题, 可以将基于机器学习的方法和基于规则的方法集成起来克服基于规则的抽取方法的这些缺点: ① 缺乏自动修正能力, 将机器学习技术应用于基于规则的系统可以提高其自动修正能力. ② 难以处理大量的特征. 当需要对大量的特征进行处理时, 通常难以预见到所有情形, 这就有可能忽略了一些有价值的特征, 而导致系统的整体性能的下降. 一些机器学习方法, 如 SVM 具有很好的可扩展性, 并能够处理高维特征空间, 将机器学习技术集成于基于规则的系统可以提高系统的整体性能. ③ 对阈值定义的随意性. 在基于规则的系统中, 阈值的定义往往比较随意.

3.4 待抽取的元数据类型问题

在现有的研究中, 不同的研究者出于应用的不同, 抽取的元数据种类和数量有所不同. 例如对于头部元数据抽取研究, Seymore 等、Han 等、McCallum 等、Fuchun Peng 等、张铭等选用了 Title、auhtor、affiliation、address 等 15 个头部元数据, 李朝光等、欧阳辉等选用了 Title、author、affiliation、keywords 等 6 个元数据; 对于引文元数据抽取研究, Eli Cortez 等选用了 Author、title、journal、Date 等 10 个元数据, Takashi Okada 等, Fuchun Peng 等选用了 Author、title、joural、Volume 等 9 个元数据. 同此看出在元数据抽取研究中, 研究者对于元数据的选择差别很大, 这在某种程度中不利于这一研究领域内研究的比较和规范. 建议学者在今后的研究中可以根据某些标准将所要抽取的元数据的种类和数量固定下来, 以便于对元数据自动抽取的研究进行规范.

3.5 测试数据集的选择

由于目前在元数据抽取领域没有一个较标准测试数据集, 所以对于抽取结果的评测和比较存在着一定的困难. 对于科技文献头部的元数据抽取测试数据集的选择, 一些研究者(如 Kristie Seymore 等、Hui Han 等、Fuchun Peng 等、张铭等)采用的是 Symore 的数据集(即将 935 篇论文头的前 500 篇作为训练集, 后 435 篇作为测试集), 而大多数研究者所采用的数据集则是自己收集. 对于科技尾部的引文元数据抽取测试数据集的选择, 一些研究者采用的是美国 CMU 大学 Cora 搜索引擎研制组提供的数据集(如 Fuchun Peng 等), 而大部分研究者的数据集则是自己构建的.

3.6 其他语种科技文献元数据的抽取

从理论和实验的研究角度来看, 目前, 国内外学者大多是对英文科技文献元数据抽取进行研究, 而对于其他语种(例如汉语、日语等)科技文献元数据抽取的研究则较少. 从较成熟的自动抽取工具的角度来看, 现有的元数据自动抽取工具都是由国外的组织和机构研制和开发的, 对英文科技文献能有效抽取, 而对于

其他语种的科技文献的抽取效果则不太理想.

参考文献

1 Wei W, King I, Lee JHM. Bibliographic attributes extraction with layer-upon-layer tagging. Proc of the ICDAR'07. Curitiba, 2007: 804–808.

2 Besagni D, Belaid A, Benet N. A segmentation method for bibliographic references by contextual tagging of fields. Proc. of the ICDAR'03. Edinburgh, 2003: 384–388.

3 李朝光,张铭,邓志鸿,杨冬青,唐世渭.论文元数据信息的自动抽取.计算机工程与应用,2002,21(10):189–191,235.

4 张铭,邓志鸿,陈捷,杨冬青,唐世渭.数字图书馆科技文献知识导航.计算机工程与应用,2002,17(1):1–3.

5 陈俊林,张文德.基于 XSLT 的 PDF 论文元数据的优化抽取.现代图书情报技术,2007,2:18–23.

6 Ding Y, Chowdhury G, Foo S. Template mining for the extraction of citation from digital documents. Proc. of the Second Asian Digital Library Conference. Taiwan, 1999: 47–62.

7 Day MY, Tsai RTH, Sung CL, Hsieh CC, Lee CW, Wu SH, Wu KP, Ong CS, Hsu WL. Reference Metadata Extraction Using a Hierarchical Knowledge representation framework. Decision Support Systems, 2007,43:152–167.

8 Eli C, da Silva AS, Marcos AG, Filipe M, de Moura ES. FLUX-CIM: flexible unsupervised extraction of citation metadata. Proc. of the JCDL'07. New York: ACM Press, 2007: 215–224.

9 Chen CC, Yang KH, Kao HY, Ho JM. BibPro: A citation parser based on sequence alignment techniques. Proc. of the IEEE AINA'08. Okinawa, Japan, 2008: 1175–1180.

10 郭志鑫,金海,陈汉华.SemreX 中基于语义的文档参考文献元数据信息提取.计算机研究与发展,2006,43(8):1368–1374.

11 高良才,汤帜,陶欣,房婧.一种自动发现、分割与标注引文元数据的方法.北京大学学报(自然科学版),2010,46(6):893–900.

12 Seymore K, McCallum A, Rosenfeld R. Learning hidden Markov model structure for information extraction. AAAI-99 Workshop on Machine Learning for Information Extraction. 1999: 37–42.

13 Geng JF, Yang J. AutoBib:Automatic extraction of bibliographic information on the Web. Proc. of the International Database Engineering and Applications Symposium (IDEAS'04). 2004.

14 Han H, Giles CL, Manavoglu E, Zha H, Zhang Z, Fox EA. Automatic document metadata extraction using support vector machines. Proc. of the 3rd ACM/IEEE－CS Joint Conference on Digital Libraries. 2003: 37–48.

15 欧阳辉,禄乐滨.基于 SVM 的论文元数据抽取方法研究.电子设计工程,2010,18(5):4–7.

16 Takasu A. Bibliographic attribute extraction from erroneous references based on a statistical model. Proc. of the 3rd ACM/IEEE－CS Joint Conference on Digital Libraries. 2003: 49–60.

17 Okada T, Takasu A, Adachi J. Bibliographic Component Extraction Using Support Vector Machines and Hidden Markov Models. Proc. of the European Conf. on Research and Advanced Technology for Digital Libraries, 2004: 501–512.

18 张铭,银平,邓志鸿,杨冬青.SVM+BiHMM:基于统计方法的元数据抽取混合模型.软件学报,2008,19(2):358–368.

19 Peng F, McCallum A. Accurate information extraction from research papers using conditional random fields. Proc. of Human Language Technology Conference and North American Chapter of the Association for Computational Linguistics. 2004: 329–336.

20 Yu J, Fan X. Metadata extraction from Chinese research papers based on conditional random fields. Proc. of the FSKD'07. Haikou, 2007: 497–501.

21 ERPANET. Packaged Object Ingest Project. http://www.erpanet.org/events/2003/rome/presentations/ross rusbridge pres.pdf.

22 National Archives. DROID(Digital Object Identification). http://www.nationalarchives.gov.uk/aboutapps/pronom/droid.htm.

23 National Library of New Zealand,Metadata Extraction Tool, http://www.natlib.govt.nz/en/what's new/4initiatives.html#extraction.

24 DC-dot.Dublin Core metadata editor.http://www.ukoln.ac.uk/metadata/dcdot/.

25 Automatic Metadata Generation,http://www.cs.kuleuven.ac.be/~hmdb/amg/documentation.php.

26 Catalogue PRO.http://ppeccatte.karefil.com/software/catalogue/catalogueDK.htm/.

27 Goodrum A, McCain K, Lawrence S, Giles C. Scholarly publishing in the Internet age: a citation analysis of computer science literature. Information Processing & Management, 2001,37:661–675.