科技论文中学术信息的提取方法综述*

胡志刚[1,2]，田文灿[1,2]，孙太安[1,2]，侯海燕[1,2]
（1.大连理工大学科学学与科技管理研究所，大连 116024；2.大连理工大学WISE实验室，大连 116024）

摘要：为更好地利用和挖掘学术论文文本，识别并提取学术论文中的学术信息已成为一种非常迫切的现实需求，在文本挖掘、信息检索、主题监测、信息计量学等领域都有广阔的应用前景。学术信息可以分为题录信息、章节信息、引文信息、引用信息和其他信息。本文综述了在PDF和HTML/XML两种不同格式的学术论文全文中，提取各类学术信息的主要方法，并指出这些方法主要面向的格式文本以及可用来提取的信息种类。最后，本文列出了提取学术信息的常用工具。

关键词：学术信息；论文全文本；信息提取；机器学习
中图分类号：G203
DOI：10.3772/j.issn.1673-2286.2017.10.007

1 引言

20世纪90年代以来，随着学术论文电子化的出现和兴起，信息技术和情报学领域的学者基于学术论文的全文本开展大量的研究工作，在信息检索[1-2]、数字图书馆[3]、主题监测和追踪[4]、自动生成摘要[5]、全文引文分析[6-9]等领域都有广泛应用。而随着开放获取运动的兴起，学术论文全文本数据的批量获取变得越来越易得[10]，为此类研究提供了更为便利的数据基础和更加广阔的应用前景。

学术论文全文本数据是文本挖掘和文献计量学研究的重要基础，包含多种学术信息，可以满足不同研究目的和功能的需要。除题录中包含的论文题目、作者、机构、期刊、期卷号等信息外，还包括章节结构、引文信息、引用信息、图表和致谢等。

近年来，面向论文全文本的学术信息提取，借助文本挖掘、自然语言处理、信息可视化、潜在语义分析、主题模型、机器学习等诸多技术手段，取得了丰富的研究成果。尤其是对元数据和引文数据的提取和解析，目前已经开发了多种高准确性和使用率的信息提取工具。

为梳理这一领域的研究进展，本文从学术论文全文的数据存储格式和学术信息的分类出发，综述PDF和HTML/XML格式中全文各类学术信息的提取方法，包括题录信息、章节信息、引文信息和引用信息。最后，本文还介绍了一些常用的学术信息提取工具或工具包。

2 学术论文文本格式的演变

随着电子计算机和互联网技术的发展，纸质学术期刊的出版模式在过去三十年日渐式微，学术期刊和学术论文的出版逐渐进入数字化时代。在学术文本数字化进程中，由Adobe公司开发的PDF文件格式凭借其优良的设计，在与DjVu、Envoy、Common Ground Digital Paper、XPS及PostScript格式的竞争中脱颖而出，成为学术出版中最受欢迎的文档格式[11]。世界知名的科技期刊出版商大多提供了PDF全文下载平台，如国外Elsevier的ScienceDirect、Springer的SpringerLink 以及John & Wiley的OnlineLibrary等；在国内的万方、维普等中文期刊全文数据库中，PDF文档也是重要的全文下载格式。

但是，PDF格式文本是一种固定版面的文本格式，不易标记学术论文中的元数据和结构信息，而HTML/XML语言正好弥补了PDF格式的这一缺陷。相比PDF格式，HTML/XML在结构化信息存储方面更加方便，可通过丰富的内部链接和外部链接，清晰地展示论文

* 本研究得到国家自然科学基金项目“开放获取背景下的全文引文分析方法与应用研究”（编号：71503031）资助。

的属性信息（如作者信息、期刊名称、卷期号等），章节结构、图表信息等；此外，这一格式还具有广阔的扩展空间，如读者通过集成引文链接服务网站（如crossref.com、dx.doi.org等），可方便地跳转到论文的参考文献页面[12]，从而极大地提高学术论文的交互性。

HTML主要用于学术论文的前台显示，而XML则主要作为学术出版工作后台的数据交换格式[13]。XML是一种标记语言，它将文档分为许多元素，并对这些元素加以标识。与HTML不同，XML将文本外观从数据的内容和结构中分离，使操作流程变得更简洁。

在XML中，元素的类型、属性等由文档类型定义（Document Type Definition，DTD）或XML Schema进行声明和定义。DTD专门用于定义文档的结构和语法，XML Schema则用于定义管理信息等更强大、更丰富的特征。在XML出现后，相继衍生出多种不同语言，包括MathML、SVG、RDF、ONIX、ePub、XHTML等。

HTML/XML格式的结构性使得它可以很好地被用于表示越来越高度结构化的学术论文。Springer、Elsevier和Wiley提供或部分提供HTML格式的全文阅读。Elsevier的ConSyn数据平台，可以直接提供XML格式全文数据的批量下载。此外，生物医学数据库PubMed Central，以及开放获取期刊PLOS、PeerJ、Frontier等都支持XML格式论文的全文下载。

总体来看，目前可供解析的学术论文全文本数据格式大致分为3种：（1）PDF格式，对读者友好，但对计算机不友好；（2）HTML格式，对读者和计算机都比较友好；（3）XML格式，对读者不友好，但对计算机友好。目前，国外的论文全文数据库，尤其是开放获取数据库，部分已升级，可同时支持3种格式的全文下载；而国内的论文全文数据库，基本上仅支持PDF格式的全文下载。

3 全文本中的学术信息

学术论文是一种具有规范结构和格式的文本。学术信息指学术论文中包含的基本元素和结构性信息，主要包括题录信息、章节信息、引文信息、引用信息，以及图表、致谢等其他信息。学术信息一般具有相对一致的位置特征、固定的模板、统一的标示性格式。

3.1 题录信息

学术论文的题录信息，一般出现在论文的开头，因此又称为论文的头信息。学术论文的元数据指由标题、作者、期刊、期卷号、DOI等题录信息构成的集合。在PDF格式的文档中，论文的标题和作者信息等一般出现在文档首页上方，用与正文不同的字体和行距识别，并常居中对齐。而在XML格式的文档中，则在论文的“Header”部分以不同的标签逐一列出。

3.2 章节信息

章节信息指学术论文各章节的标题、位置、边界等信息。学术论文一般都具有约定俗成的逻辑结构，如自然科学领域最常见的IMR&D四节式结构[14]。其中，引言（introduction）部分主要描述研究背景和提出研究问题；方法（method）部分主要给出研究所使用的实验材料及其实现方法、数据及分析算法等；结果（result）部分主要展现研究结果，该部分一般包含丰富的图表信息；而结论（conclusion）部分总结论文作者的成果和贡献，回答论文开始提出的问题。

在HTML/XML格式的文档中，论文的章节信息通常以“<section>”标签直接标出。而在PDF格式的文档中，章节信息可以基于字体的大小、是否加粗以及是否留白等进行识别。

3.3 引文信息

在学术论文的正文之后，大多会列出论文中所有引用的文献信息。引文信息包含参考文献的作者、标题、出版物名称、发表年份、卷期号、页码等信息，部分还标有DOI标识符。不同出版商或不同学会出版的期刊，通常都有其特有的引文样式，国外期刊常用的样式有APA 6.0（美国心理学学会样式）、CMS（芝加哥样式）、MLA（美国现代语言协会样式）等，而国内期刊主要遵循的标准是GB/T 7714—2005。目前，在引用样式网站上列出的引文样式已达8 000多种。

在PDF文档中，引文信息通常位于正文的末尾，逐行列出，且每条引文都应遵循标准的固定样式。对于XML格式的文档来说，引文信息是在文档的最后逐条逐项列出。

3.4 引用信息

引用信息与引文信息不同，指引文在施引文献正文

中被引用的位置和语境。引文在正文中被引用时的标识，大致可分成两种：一种是大多数英文期刊所采用的标记；另一种是大部分中文期刊和部分英文期刊所采用的数字序号，通常在引用位置以上角标的样式标出。两种标记方法所对应的参考引文列表的排列方式不同，前者通常按照字母顺序进行罗列，后者通常按照引用位置出现的顺序进行罗列。

在PDF格式论文中，引用信息可以通过特殊样式来识别。对XML格式的全文数据来说，由于引用会以专门的标签给出，因此引用信息的识别要相对简单得多。

目前在信息检索领域已广泛地开展题录数据和引文信息提取，相对来说，对引用信息的提取还比较少。引用位置、强度和语境，可以用来甄别引用功能、动机和情感，随着引用内容分析的兴起，引用信息的提取和应用将变得越来越普遍。

3.5 其他信息

除论文的题录信息、章节信息、引文信息和引用信息外，学术论文中可供提取和分析的学术信息还包括摘要信息（尤其是结构化摘要）、图信息[15]、表信息[16]、公式[17]、致谢[18]（通常含有基金资助信息）和附录信息等。这些信息在学术论文的话语分析[19]、论文抄袭检测研究等方面，也具有重要的价值和意义。

4 学术信息提取方法

学术信息是论文的结构化要素，既可以作为编制文献数据库的索引要素，又可以作为文献计量研究的分析要素。对学术信息的识别和提取，是信息检索、文献管理、文献存储和文献计量研究的基础和前提，不论是CiteSeerX、Google Scholar等文献数据库，还是Mendeley、Zotero等文献管理软件，都离不开基于全文本的学术信息提取工作。

对于学术论文中学术信息的提取，按照问题的复杂程度和提取方法的难易程度，可以分为基于模板、基于规则和基于机器学习三种提取方法。

基于规则和基于机器学习的提取方法，主要针对PDF文档，在这种格式的文档中，论文的学术信息没有被直接标出，需要通过论文的排版、位置、格式等规则，或综合借助特征词典对学术信息进行识别和分类。而基于模板的提取方法，主要针对XML/HTML格式的学术论文，由于XML/HTML格式全文对论文中的学术信息进行格式化标注，因此，只需要基于XML文档的DTD识别对应的标签，就可以提取出所需的学术信息。

此外，由于提取学术信息的对象不同，所采用的提取方法也不同。如对于引文信息的提取一般采用基于模板或基于规则的方法，因为引文有固定的样式和模板，掌握引文书写的规则很容易利用其来反推引文的各个组成部分。而对于题录信息、引用信息这种非结构化内容的提取，则需要更多地借助机器学习进行识别和提取。

4.1 基于模板的提取方法

对于XML格式的结构化文本来说，学术信息的提取相对简单。在XML格式的全文中，结构化学术信息都以标签进行标记，而且所用标签的含义在DTD中进行规定和说明，因此基于模板的提取方法相对简单且直接。

目前，在学术期刊界，利用最广泛的XML框架和标签集是期刊论文标签集（Journal Article Tag Suite，JATS），这一标准最早由美国国家医学图书馆开发，于2012年被确立为美国国家标准（NISO Z39.96）。JATS定义了XML文件中的元素和元素的属性、排列方式、包含内容等，在JATS中共有246个元素和134种属性。

基于模板的学术信息提取，主要通过对XML全文的解析函数完成。一些常见的程序语言（如PHP、Java、Python等）中基本都含有对XML的解析函数或命令，调用这些函数或命令，即可将XML文件中的元素信息提取到数组，方便用户进一步存放到数据库和数据表中。

相对于基于规则和基于机器学习的方法，基于模板的方法具有更高的准确度。但是由于其完全依赖文档的框架和标签集，因此对于某些质量不高的XML格式数据，可能出现提取失败或中断的状况。

该方法可以广泛应用于文档中题录信息、章节信息、引文信息、引用信息和其他类型信息的提取。基于模板的学术信息提取的代表性工具有ParaCite、InfoMap等。Flynn等利用模板和字符串查找函数，并提取学术论文和研究报告中的元数据[20]。胡志刚等设计一种在XML格式全文中提取引用位置、引用语境信息的全文引文分析系统[21]。

4.2 基于规则的提取方法

基于规则的提取方法是基于一系列事先定义好的

规则和流程，对论文的题录信息、引文信息或引用信息等各类学术信息进行提取。该提取方法的背景和前提是学术论文通常会遵从一定的结构和格式。学术信息提取的规则设计可以基于知识、经验和启发式方法，因此基于规则的提取方法又称基于知识的提取方法。可以利用如下规则来提取论文的题录信息：（1）标题通常位于正文的开头且在全篇中字体最大；（2）作者位于标题的下方；（3）各作者名称的字体相同；（4）机构位于作者名称下方；（5）机构的字体相同；（6）如果只有一个机构，那么所有作者都属于这个机构；（7）章节标题比正文字体大。

基于规则的提取方法主要针对PDF、HTML或其他富文本格式，其准确率一般低于基于模板的方法，高于基于机器学习的方法。但是这种方法费时费力，尤其是规则较多时，该方法在题录信息、引文信息的提取中应用较广，很多常用的工具（如CiteSeerX[22]、Google Scholar[23]等）都基于或部分基于这种方法。Giuffrida等曾利用基于规则的方法提取PostScript这种半结构化全文中的题录信息和章节信息[24]；Groza等则提出一种面向PDF格式文档首页的元数据提取规则，主要基于字体的格式或位置进行提取，具有较高的准确性[25]。

4.3 基于机器学习的提取方法

基于模板或基于规则的方法，非常依赖专家的规则方案和文本的规范程度，一旦文本格式的复杂程度超出专家可以进行规则化的范围，就必须依赖机器学习的方式进行提取。

机器学习方法是通过对训练数据的学习获得信息抽取的模式，并对未知数据进行判定和预测，主要用于元数据、引文信息和引用信息等相对复杂信息的抽取任务。如从引文中识别作者、期刊名、标题等信息，或者对引用语境进行情感分析、实体标注等。对于某些难以利用模板或规则进行抽取的不规范文本，基于机器学习的方法不失为一种行之有效的选择。

4.3.1 基于支持向量机的方法

支持向量机（Support Vector Machine，SVM）是机器学习中的一种基本分类算法。SVM应用于文本和超文本的分类，可显著减少所需训练样本数，提高分类效率。在对学术信息进行抽取时，需先将抽取问题转换成二元分类问题。如将抽取作者信息的问题，转换成判断一个字符串是否为作者的问题。

SVM方法的基础，是对要分类的字符串进行特征选择，并赋予不同权重。一般来说，学术信息相关的文本主要由以下特征组成：（1）格式特征，如首字母是否大写、是否包含数字、是否包含简写、是否符合邮箱的正则式等；（2）位置特征，相关文本在句首、句中还是句尾；（3）词典特征，如是否可以匹配姓名词典、机构词典、时间词典等。SVM的目标，就是从这些特征中选择对于分类真正重要的特征，以及确定间隔平面的特征向量（称为支持向量）。

在学术信息的抽取任务中，SVM既可用于题录信息的标注[26]，又可用于引文信息的解析[27]。在对引文进行解析时，主要考虑的特征包括：是否存在于作者词典中，是否存在于期刊词典中，是否包含“et al”，是否包含“pp.”或“p.”，是否全为数字，是否全为字母，以及在整个句子中的位置等。

基于SVM分类方法的缺点是只能根据文本块自身的特征进行分类。对于学术信息的抽取来说，各文本块间的文法和语法规则（如各文本块出现的顺序，文本块间的分隔词或字符等），对判断字符串的类型是非常重要的，其重要程度有时甚至超过文本块自身的内容。因此，SVM在准确度上一般明显低于专门面向序列标注的隐马尔可夫模型（Hidden Markov Model，HMM）和条件随机场（Conditional Random Fields，CRF）。

4.3.2 基于HMM的方法

自然语言本质上可以看作一种由词语组成的序列，词语间不是彼此孤立的，在前后顺序和关联上需要遵守一定的文法和语法规则。HMM的方法和CRF是自然语言处理时常用的两种数学模型，可以有效处理序列数据的标注问题，因此也被大量应用到学术信息的抽取过程中。

HMM是19世纪60年代由Baum等提出的[28]，1980年，贝尔实验室的Rabiner等对HMM进行简化，并率先在语音识别中运用和推广[29]，此后被广泛应用于语音识别、实体识别、词性标注、信息抽取等领域。

HMM分别描述了一个可观察的和一个隐性的随机过程，隐性状态间的转换过程对应一个转移概率矩阵，需要借助可观察的随机过程进行推断。对于学术文本来说，作者、期刊名、标题、机构、期卷号等学术信息类型就

是一个隐性的状态序列。学术信息抽取的过程，就是标记引文中的各个部分所属的状态（即学术信息类型）。

HMM在学术文本领域的应用在20世纪90年代开始出现，早期主要集中应用在学术论文中的实体识别（如识别医学论文中的症状、药物、基因等），随后扩展到对引文信息的解析[30-32]和头信息的识别[33-37]。

HMM具有易于建立、不需要大规模的词典集与规则集、适应性好和精度较高等优点。在学术信息提取中，如果通过人工制定的规则难以达到较好的提取效果，就可以考虑采用HMM来处理该问题。

4.3.3 基于CRF的方法

CRF是另一种广泛使用的序列标注模型，由Lafferty等提出[38]。HMM依赖“状态转移过程中当前状态只与前一状态有关”这一个局部性假设，而CRF具有表达元素长距离依赖性和交叠性特征的能力，更易于处理关联较强的信息抽取工作，如对于引用语境信息的抽取[39]。在抽取效果方面，CRF也展现了优于HMM的提取效果[38,40-41]。

与SVM和HMM一样，CRF在进行学术信息抽取时也依赖特征的选则与抽取，而且CRF可以为特征集中的各种特征赋予不同权重。实验表明，位置和序列信息在特征空间中的权重越大，抽取的效果越好[40]。

由于具有更高的准确率和召回率，近年来，CRF已经成为学术信息领域用得最多的一种模型。尤其是借助经典的CRF++工具包，基于CRF算法的学术抽取工具的开发变得更加方便，如对于引用语境信息的抽取。

5 学术信息提取工具

本文主要分析7个常用的学术信息提取工具。这些工具大部分提供开源下载或在线服务界面，用户可以借助这些工具从XML、PDF或纯文本中提取论文的题录信息、引文信息等。

从开发目的来看，有的工具只面向单一类型信息的提取，如ParaCite只能解析引文信息；而有些则提供“一揽子”学术信息的提取功能，如ParsCit程序中同时集成了题录信息、章节信息和引文信息的提取。从应用渠道来看，有些工具用于文献管理，如Mendeley，Zotero等；有些工具则用于构建数字图书馆或文献搜索引擎，如CiteSeerX、Google Scholar等。

5.1 CiteSeerX

CiteSeerX的前身为CiteSeer，是由美国普林斯顿大学NEC研究院研制开发的一款最先利用自动引文索引技术建立的科学文献数据库和搜索引擎[22,42]。在1997年，CiteSeer就开始自动爬取互联网上Postscript和PDF格式的开放获取学术论文，比Google Scholar早7年。2007年，研发人员在对原系统运行中暴露的问题和用户反馈意见进行分析的基础上，为该搜索引擎重新设计系统结构和数据模型，并改名为CiteSeerX。

作为数字图书馆，目前CiteSeerX中可检索到的论文数量超过700万篇，主要涉及计算机科学领域。作为搜索引擎，CiteSeerX系统的功能主要包括：（1）利用主题词、作者等检索文献；（2）检索结果会列出检索文献的题录信息、引用语境，某一具体文献的施引文献、参考文献以及相关文献（共被引文献）等。

CiteSeer向用户免费提供文献学术信息提取的应用程序接口——CiteSeer Extractor，为用户提供PDF格式的学术论文解析和学术信息提取服务。用户通过CiteSeer Extractor可以从PDF文档中提取元数据、引文信息和正文章节，并将提取的结果以XML、JSON或BibTex等格式返回给用户。

CiteSeer Extractor，集成了一系列的开源工具包（包括ParsCit、SVM Header Parse、PDFBox等），这些工具包被用来完成自动引文索引、自动元数据提取、引文统计、生成引文链接、作者消歧、引用语境提取等一系列工作。CiteSeer Extractor也是开源的，用户可以免费获取源代码并进行修改。

5.2 Mendeley

Mendeley是一个免费的参考文献管理工具与学术社交媒体[43]，2008年被推出，凭借其超前的理念和强大的产品功能获得多项欧洲大奖，2013年被Elsevier公司收购。Mendeley可以帮助使用者管理和组织学术文献，可以在线与其他研究者合作交流，以及发现最新研究成果。

对PDF文档的解析和学术信息的自动提取是Mendeley区别于其他软件的最大特色，它内置了PDF阅读器，可以方便浏览和标注全文，并支持对PDF全文的检索。更重要的是，它可以轻松解析用户导入的PDF全文数据，提取出其中的题录信息、章节框架，以便更有

效地管理文献。

Mendeley对PDF的解析基于Grobid开源程序包，其提取学术信息的具体步骤为：（1）利用pdf2xml程序将PDF转换成带格式（包括大小、字体和位置）的文本文件；（2）将文本中的各类信息转换成分类器所需要的特征，然后利用开源工具包Grobid的元数据提取程序包对论文中的题目、作者、摘要等信息进行提取；（3）利用提取的学术信息生成一个检索式并提交，以便与Mendeley、Arxiv、PubMed和CrossRef等数据库中现有文献进行比较，从而进一步丰富论文的元数据信息。

与Mendeley类似的工具还包括Zotero、Docear、PDFmeat等，这些工具中同样集成了PDF文献的学术信息提取功能。其中Docear是由德国马格德堡大学的SciPlore研究小组开发的一款兼具文献管理和论文写作功能的思维导图软件[44]，它可以从导入的PDF文献中提取题录信息和章节信息。

5.3 ParsCit

ParsCit是由靳民彦等开发的一个功能齐全、性能强大的学术信息提取工具，它既可以进行引文信息和引用信息提取，也可以对论文的题录信息和章节信息提取[45]。其中，对题录信息和章节信息的提取由与ParsCit同源的另外两个程序ParsHead和SectLabel协助实现。

ParsCit是一种基于CRF的信息提取工具。其代码是开源的，并且代码中包含训练集、特征生成器等。ParsCit的安装和运行需要ruby、perl和CRF++嵌套包，不过，ParsCit提供在线提取功能，支持对TXT、PDF、XML等格式的论文进行在线解析和提取。其中，XML格式的文档默认支持OmniPage的DTD框架。

5.4 ParaCite

ParaCite是机构知识库服务商Eprints开发的一个引文信息提取工具和检索平台[46]，集成在Eprints的软件系统中，用于对引文的解析（reference parser模块）和引用文献的检索（reference resolver模块）。在引文的解析方面，ParaCite利用基于模板匹配的方法，将引文字符串与设定的引文模板集（目前包含235个常用模板）逐一进行匹配，找到与待解析引文最符合的模板并据此将引文切分为作者、年份、题目、期刊名、期卷号等信息单元。在引用文献的检索方面，ParaCite提供被引论文的检索界面，用户可将引文字符串输入到检索框中进行检索，ParaCite对输入的引文字符串进行解析，分别生成其在Google Scholar、CiteBase、Google和CiteSeer等数据库中的openurl，并发送给各数据库等待返回的检索结果。

5.5 GROBID

文献目录数据生成器（GeneRation Of BIbliographic Data，GROBID）是一款基于CRF算法的学术信息提取工具[47]，用于在PDF格式的科技文献中提取、解析学术信息，并进行TEI编码的结构化存储。GROBID利用Java语言进行开发，并集成其他开源程序包。首先利用Xpdf程序对PDF进行预处理，然后在学术信息提取时通过JNI调用法国LIMSI-CNRS实验室开发的Wapiti CRF Library程序包。

GROBID功能强大，可以提供55种学术信息的提取和识别。CROBID程序包中包含批处理程序、基于网络的RESTful API、Java API、相对通用的评价框架和半自动生成的训练级数据，面向用户开源。在学术信息的解析和提取方面，GROBID具有很高的准确度和运行效率。从程序开发者基于MacBook Pro的测试结果来看，平均每秒可以完成3篇PDF文档的解析和提取，并且在18秒内完成3 000条引文的解析。

由于其卓越的性能，GROBID在很多文献数据库和存储平台中有大量应用，其中包括ResearchGate、Mendeley、HAL Research Archive、the European Patent Office、INIST和CERN等。用户还可以通过其官网在线使用GROBID的解析服务。

5.6 PDFx

PDFx是由奥地利程序开发者Hager利用Python开发的一款学术信息提取和参考文献下载工具。Hager指出其开发这一工具的背景和初衷是，当读者读到一篇不错的论文时，往往想要下载这篇论文中的所有参考文献，但这通常是一件非常麻烦的事情，尤其当参考文献较多时，下载就更加费时费力。其所开发的PDFx工具可以提取PDF文献中的参考文献、元数据和正文文本，下载这些参考文献的PDF（需要用户所在机构购买相应的全文数据库）。PDFx同样面向用户开源。

5.7 INFOMAP

INFOMAP是Day等开发的一款基于本体知识表示的引文信息提取工具[48]，可以提取引文中的作者、标题、期刊和期卷号等信息。本体是一个形式化的、共享的、明确化的、概念化的规范，用本体表示知识的目的是统一应用领域的概念，并构建本体层级体系表示概念间语义关系，实现人类、计算机对知识的共享和重用。本体层级体系的基本组成部分是五个基本的建模元语，分别为类、关系、函数、公理和实例。领域本体知识库中的知识，不仅通过纵向类属分类，而且通过本体的语义关联进行组织和关联，再利用这些知识进行推理，从而提高学术信息识别准确率。

基于对APA、IEEE、ACM、BIOI、JCB、MISQ的引文数据集所做的实验，结果表明，INFOMAP的准确度平均高达97.87%[49]。

6 结论

学术信息提取方法和工具的大量出现，标志着对学术论文的全文分析正日趋成熟。按照Shneider对于学科领域的四阶段划分[50]，一个研究领域的发展可以分成四个阶段：I-研究对象和有关概念的形成阶段，II-大量方法和工具的开发阶段，III-研究问题调研和解答阶段，IV-隐性知识的显性化阶段。其中，阶段II-大量方法和工具的开发，是从理论到应用、从概念到实践的必经之路。显然，当前全文分析正处在一个非常关键的阶段。

本文综述了论文全文本中提取学术信息的主要方法和主要工具。随着开放获取运动的兴起，学术论文全文本数据的批量获取变得越来越易得，在此背景下，准确高效地提取全文数据中的学术信息已经成为重要的热点课题。

学术信息的提取具有重要的学术价值和应用前景。在学术论文的全文本中提取学术信息，不仅可以有效地提高信息检索的功能和精度，更好地为用户提供知识服务，而且通过对学术信息的计量和统计，可以更好地基于学术论文全文进行深度的知识挖掘和知识发现。

参考文献

[1] MAYR P,SCHARNHORST A.Combining bibliometrics and information retrieval:preface[J].Scientometrics,2015,102(3):2191-2192.

[2] LIU S,CHEN C,DING K,et al.Literature retrieval based on citation context[J].Scientometrics,2014,101(2):1293-1307.

[3] WILLIAMS K,WU J,CHOUDHURY S R,et al.Scholarly big data information extraction and integration in the CiteSeer x digital library[C]//IEEE International Conference on Date Engineeri Workshops.[S.l.]:[s.n.],2014:68-73.

[4] WANG X,CHENG Q,LU W.Analyzing evolution of research topics with NEViewer: a new method based on dynamic co-word networks[J].Scientometrics,2014,101(2):1253-1271.

[5] YE S,CHUA T S S,KAN M Y,et al.Document concept lattice for text understanding and summarization[J].Information Processing and Management,2007,43(6):1643-1662.

[6] LIU X,ZHANG J,GUO C.Full-text citation analysis:a new method to enhance scholarly networks[J].Journal of the American Society for Information Science and Technology,2013,64(9):1852-1863.

[7] GLENISSON P, GLÄNZEL W,PERSSON O.Combining full-text analysis and bibliometric indicators.A pilot study[J].Scientometrics, 2005,63(1):163-180.

[8] 赵蓉英,曾宪琴,陈必坤.全文本引文分析——引文分析的新发展[J].图书情报工作,2014,58(9):129-135.

[9] 胡志刚.全文引文分析:理论、方法与应用[M].北京:科学出版社,2016.

[10] 胡志刚,侯海燕,林歌歌.从书信沙龙到开放获取——刍议学术学术论文形态的演化[J].数字图书馆论坛,2016(10):32-37.

[11] 张立.数字出版相关概念的比较分析[J].中国出版,2006(12):11-14.

[12] ZOU J,LE D,THOMA G R.Locating and parsing bibliographic references in HTML medical articles[J].International Journal on Document Analysis and Recognition,2010,13(2):107-119.

[13] 白杰,杨爱臣.XML结构化数字出版的特点与流程[J].出版广角,2015(5):28-31.

[14] SOLLACI L B,PEREIRA M G.The introduction,methods,results, and discussion(IMRAD) structure:a fifty-year survey[J].Journal of the Medical Library Association Jmla,2004,92(3):364-367.

[15] CHOUDHURY S R,TUAROB S,MITRA P,et al.A figure search engine architecture for a chemistry digital library[J].2013:369-370.

[16] LIU Y,BAI K,MITRA P,et al.TableSeer: automatic table metadata extraction and searching in digital libraries[C]//JCDL'07. Vancouver: [s.n.],2007:91-100.

[17] JIN J,HAN X,WANG Q.Mathematical Formulas Extraction[C]// International Conference on Document Analysis and Recognition, IEEE.[S.l.]:[s.n.],2003:1138-1141.

[18] COUNCILL I G,GILES C L,HAN H,et al.Automatic acknowledgement

indexing:expanding the semantics of contribution in the CiteSeer digital library[C]//International Conference on Knowledge Capture,Banff:[s.n.],2005:1-8.

[19] SARIC J,CIMIANO P.Ontology-driven discourse analysis for information extraction[J].Data & Knowledge Engineering,2005,55:59-83.

[20] FLYNN P,LI Z,MALY K,et al.Automated template-based metadata extraction architecture[C]//International Conference on Asian Digital Libraries:Looking Back 10 Year and Forging New Frontiers.[S.l.]:Springer-Verlag,2007.

[21] 胡志刚,陈超美,刘则渊,等.基于XML全文数据引文分析系统的设计与实现[J].现代图书情报技术,2012(11):71-77.

[22] GILES C L,BOLLACKER K D,LAWRENCE S.CiteSeer:an automatic citation indexing system[C]//Proceedings of the third ACM conference on Digital libraries.[S.l.]:ACM,1998:89-98.

[23] GOOGLE.Inclusion Guidelines for Webmasters:Indexing Guidelines[EB/OL].[2017-08-01].https://scholar.google.com/intl/zh-CN/scholar/inclusion.html#indexing.

[24] GIUFFRIDA G,SHEK E C,YANG J.Knowledge-based metadata extraction from PostScript files[C]//Proceedings of the 5th ACM Conference on Digital Libraries.New York:ACM Press,2000:77-84.

[25] GROZA T,HANDSCHUH S,HULPUS I.A document engineering approach to automatic extraction of shallow metadata from scientific publications[R/OL].[2017-08-01].https://www.researchgate.net/publication/237536549_A_ DOCUMENT_ENGINEERING_APPROACH_TO_AUTOMATIC_EXTRACTION_OF_SHALLOW_METADATA_FROM_SCIENTIFIC_PUBLICATIONS.

[26] HAN H,GILES C L L,MANAVOGLU E,et al.Automatic document metadata extraction using support vector machines[C]//Joint Conference on Digital Libraries.[S.l.]:IEEE,2003:37-48.

[27] ZHANG X,ZOU J,LE D X,et al.A structural SVM approach for reference parsing[J].BMC Bioinformatics,2011,12(3):1-7.

[28] BAUM L E,PETRIE T.Statistical inference for probabilistic functions of finite state Markov chains[J].Annals of Mathematical Statistics,1966,37(6):1554-1563.

[29] RABINER L R.A tutorial on hidden Markov models and selected applications in speech recognition[C]//Proceedings of the IEEE.[S.l.]:IEEE,1989,77(2):257-286.

[30] HETZNER E.A simple method for citation metadata extraction using hidden markov models[C]//Joint Conference on Digital Libraries.[S.l.]:[s.n.],2008:280-284.

[31] 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.

[32] CUI B G,CHEN X.An improved Hidden Markov Model for literature metadata extraction[C]//International Conference on Advanced Intelligent Computing Theories and Application:Intelligent Computing,[S.l.]:Springer Berlin Heidelberg,2010,6251(4):205-212.

[33] PARK D C,HUONG V T L,WOO D M,et al.Information extraction system based on Hidden Markov Model[M].Berlin:Springer Berlin Heidelberg,2009:52-59.

[34] SONG M,SONG I Y,HU X H,et al.KXtractor:an effective biomedical information extraction technique based on mixture Hidden Markov models[M].Berlin:Springer Berlin Heidelberg,2005:68-81.

[35] ZHONG P,CHEN J,COOK T.Web information extraction using generalized Hidden Markov Model[C]//IEEE Workshop on Hot Topics in Web Systems and Techologies.[S.l.]:IEEE,2006:1-8.

[36] XIAO J,ZOU L,LI C.Optimization of Hidden Markov Model by a genetic algorithm for web information extraction[J].International Journal of Computational Intelligence Systems,2007.

[37] CHI C Y,ZHANG Y.Information extraction from Chinese papers based on Hidden Markov Model[J].Advanced Materials Research,2014:846-847,1291-1294.

[38] LAFFERTY J D,MCCALLUM A,PEREIRA F C N.Conditional random fields:probabilistic models for segmenting and labeling sequence data[C]//Proceedings of the 18th International Conference on Machine Learning.[S.l.]:[s.n.],2001,3(2):282-289.

[39] SCHWARTZ A S,DIVOLI A,HEARST M A.Multiple alignment of citation sentences with conditional random fields and posterior decoding example of unaligned citances[J].Computational Linguistics,2007(6):847-857.

[40] PENG F,MCCALLUM A.Information extraction from research papers using conditional random fields[J].Information Processing and Management,2006,42(4):963-979.

[41] PINTO D,MCCALLUM A,WEI X,et al.Table extraction using conditional random fields[C]//Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval.[S.l.]:[s.n.],2003:235-242.

[42] ORORBIAII A G,WU J,KHABSA M,et al.Big scholarly data in CiteSeerX:information extraction from the web[C]//International Conference.[S.l.]:[s.n.],2015:597-602.

[43] HENNING V,REICHELT J.Mendeley-A Last.fm for research?[C]//IEEE 4th International Conference on Escience.[S.l.]:IEEE,2008:327-328.

[44] BEEL J,GIPP B,LANGER S,et al.Docear:an academic literature

suite for searching,organizing and creating academic literature[C]// Proceedings of the 11th Annual.[S.l.]:[s.n.],2011:4-6.

[45] COUNCILL I G,GILES C L,KAN M Y.ParsCit:an open-source CRF reference string parsing package[J].LREC'08:Proceedings of the 6th International Conference on Language Resources and Evaluation.[S.l.]:[s.n.],2008(3):661-667.

[46] GUPTA D,MORRIS B,CATAPANO T,et al.A new approach towards bibliographic reference identification,parsing and inline citation matching[C]//Communications in Computer and Information Science. Berlin:Springer Berlin Heidelberg,2009,40:93-102.

[47] LOPEZ P.GROBID:combining automatic bibliographic data recognition and term extraction for scholarship publications[C]// Proceedings of the 13th European Conference on Digital Library. Corfu:[s.n.],2009:473-474.

[48] DAY M Y,TSAI T H,SUNG C L,et al.A knowledge-based approach to citation extraction[C]//Proceedings of the 2005 IEEE International Conference on Information Reuse and Integration.[S.l.]:[s.n.], 2005:50-55.

[49] CHEN C C,YANG K H,KAO H Y,et al.BibPro: a citation parser based on sequence alignment techniques[C]//22nd International Conference on Advanced Information Networking and Applications.[S.l.]:[s.n.], 2008:1175-1180.

[50] SHNEIDER A M.Four stages of a scientific discipline;four types of scientist[J].Trends in Biochemical Sciences,2009,34(5):217.

作者简介

胡志刚，男，1984年生，讲师，硕士生导师，研究方向：全文引文分析、科学计量学。
田文灿，男，1995年生，硕士研究生，研究方向：全文引文分析、科学知识图谱。
孙太安，男，1991年生，硕士研究生，研究方向：科学计量学与信息计量学。
侯海燕，女，1971年生，博士，教授，博士生导师，通讯作者，研究方向：科学学与科技管理、科学计量学，E-mail：htieshan@dlut.edu.cn。

A Method Review on Academic Information Extracting from Scientific Papers

HU ZhiGang[1,2], TIAN WenCan[1,2], SUN TaiAn[1,2], HOU HaiYan[1,2]
(1.Institute of Science of Science and Science and Technology Management, Dalian University of Technology, Dalian 116024, China;
2.WISE Laboratory, Dalian University of Technology, Dalian 116024, China)

Abstract: In order to make better use of rich information in academic papers, it is a very urgent and realistic requirement to identify and extract academic information within. The academic information extracting has a broad application prospect in text mining, information retrieval, theme monitoring, information metrology and many other fields. There are five kinds of academic information, such as title information, section information, citation information, reference information and other information. This paper reviews the methods of academic information extracting from the full text of academic papers. Different methods could be used to extract different kinds of academic information from different types of full texts, PDF or HTML/XML. Finally, the paper also lists the current tools for extracting academic information.

Keywords: Academic Information; Full Text; Information Extraction; Machine Learning

（收稿日期：2017-08-28）