

基于引文上下文的相关研究辅助生成系统设计*^{*}

王鑫^{1,2}, 程齐凯^{1,2}, 李信^{1,2}, 陆伟^{1,2}

(1. 武汉大学信息管理学院, 武汉 430072; 2. 武汉大学信息检索与知识挖掘研究所, 武汉 430072)

摘要: 本文对学术文本引文上下文的定义及其识别研究进展进行梳理, 考虑科研人员在相关研究撰写过程中或期刊编辑在审稿过程中的特定信息需求, 探讨基于引文上下文辅助生成相关研究章节的可行性。在此基础上, 对基于引文上下文的相关研究辅助生成系统的系统思路、功能模块进行设计, 并以ScienceDirect数据库在1957—2014年收录的289 926篇计算机领域的科研文献全文网页数据作为数据源, 实现相关研究辅助生成系统RWGS。结果表明, RWGS可较好满足科研人员在撰写相关研究章节或期刊编辑在审稿时更细粒度的信息需求, 对传统学术数据库的检索结果有一定优化效果, 同时将文献检索、文献阅读和相关研究写作三个过程有机地整合, 有效改善系统用户体验, 提升学术创作效率。

关键词: 相关研究; 辅助生成系统; 引文上下文; 计算机领域

中图分类号: G250.7

DOI: 10.3772/j.issn.1673-2286.2017.08.003

1 引言

科研工作者在进行学术创作时, 需要广泛收集和阅读与研究主题相关的研究成果, 以对主题的研究现状、研究热点、发展趋势和存在问题进行全面准确地把握; 并在此基础上形成相关研究章节, 以支撑其研究, 避免重复性和错误性的工作。然而, 随着学术文献的爆炸式增长与多学科合作研究的广泛开展, 传统的依靠人力来综述文献的方式越来越困难^[1]; 同时, 这也给学术期刊审稿的速度、效率和准确度带来挑战。因此, 在科学创作和审稿过程中, 针对特定研究主题, 如何快速全面地获取相关研究; 如何快速对已有研究成果形成客观全面的描述和评价; 如何利用计算机自动进行文献回顾, 并完成相关研究章节的辅助生成等问题的解决, 对提升科研人员的学术创作效率、辅助期刊编辑审稿和有效进行学术传播具有重要的现实意义。通过文献调查发现, 计算机科学、情报学和可视化等领域的学者已经

进行探索, 并开发了一系列具有实用价值的学术创作辅助系统(如文献分析系统CiteSpace^[2]、CiteRiver^[3], 辅助写作系统FLOW^[4]、WriteAhead^[5])。通过比较发现, 现有学术创作辅助系统大多将检索文献、阅读文献和论文写作三个相辅相成、互为交替的过程进行人为分割, 导致用户体验差, 实际辅助效果不佳。传统的文献检索系统仅对文献的题录信息建立索引, 导致返回的结果无法满足用户在撰写相关研究章节时的特定信息需求。此外, 现有的文献分析系统也基本上不涉及文献的引文上下文; 而当学者在撰写论文的相关章节或期刊编辑在审稿时, 很大程度上希望直接得到系统返回的结果是文献中的相关研究章节, 甚至直接是相关研究章节对特定研究成果的描述和评价语句(即引文上下文)。

基于此, 本文从学术文献引文上下文的角度出发, 构建基于引文上下文的相关研究辅助生成系统, 将检索文献、阅读文献和相关研究章节写作有机结合, 在一定程度上弥补已有研究的不足。

* 本研究得到中国博士后科学基金项目(编号: 2016M602371)和国家自然科学基金青年项目“基于深度语义挖掘的引文推荐多样化研究”(编号: 71704137)资助。

2 相关研究综述

2.1 学术创作辅助系统

随着计算机信息技术的快速发展和学术大数据时代的来临,为提升学术创作效率,计算机科学、信息科学和科学学领域的学者针对学术创作过程的不同环节,设计和开发了相应学术创作辅助系统,根据系统主要功能将其分为学术检索系统、文献分析系统和辅助写作系统。

学术检索系统主要基于数据库和关联数据技术,对科研文献的元数据建立索引,为用户提供文献检索和导航服务,并提供简单的基于元数据的文献统计分析功能,如Web of Science、ScienceDirect、PubMed、中国知网、万方数据库、维普网等。文献分析系统的主要功能是帮助用户更好地阅读和理解科研文献,这类系统通常基于文献计量理论和知识图谱技术来实现对科研文献的自动化语义分析和可视化,以使用户快速全面地把握研究主题的热点问题、整体态势和研究趋势,帮助用户阅读和理解科研文献。具有代表性的文献分析系统主要有CiteSpace^[2]、VOSViewer^[6]、NEViewer^[7]和CiteRivers^[3],CiteSpace和VOSViewer通过引文分析和可视化来向用户直观快速地展示研究领域的新兴热点、发展趋势,而NEViewer和CiteRiver分别利用桑基图和河流图等可视化技术,将文献在时间序列上的统计信息返回给用户。

辅助写作系统旨在帮助用户提高写作效率。目前已有的辅助写作系统主要分为双语写作系统、摘要生成系统和主题推荐系统,这三类系统分别从语言学、自动摘要生成和引文推荐的角度来辅助用户学术写作。Chen等为非英语母语学者开发的辅助写作系统FLOW^[4]、杨秉哲开发的摘要辅助写作系统WriteAhead^[5]、孔行通过LDA主题模型开发的例句推荐辅助写作系统^[8]都是典型的辅助写作系统。

国内外学者在提高用户检索、阅读理解和学术论文写作的效率上,已经进行大量的尝试,设计和开发了一系列学术创作辅助系统,但仍存在不足。一方面,学术检索系统仅依赖检索词进行字符匹配返回的检索结果存在大量冗余、无关的文献,无法满足科研工作者在学术写作时的特定信息需求;另一方面,现有的学术辅助创作系统大多是人为地将这个有机整体进行分割。此外,相关辅助写作系统的研究还基本处于空白。

基于以上分析结果,本文聚焦相关研究辅助生成系统的设计和实现,从引文上下文识别的视角出发,充分考虑科研工作者在撰写相关研究时的特定信息需求,有针对性地匹配检索结果,从而对传统学术数据库的检索结果进行优化,改善系统的浏览和导航体验。将文献检索、文献阅读和论文写作三者有机地结合,利用引文上下文的自由组合和用户自定义书写的半自动化写作模式,改善用户体验,提高论文写作效率。

2.2 引文上下文及其自动识别

引文上下文的研究源于科学家对传统计量视角的引文分析可靠性的怀疑,因其仅从引用的角度对施引文献和被引文献的关系进行探究,而忽视引文内容、功能、情感和重要性等语义信息。1975年,Chubin等提出引文内容分析,即以引文上下文为依据对引文的性质进行分析,深度挖掘施引文献与被引文献间的语义关系,但没有具体给出引文上下文的定义^[9];1999年,Nanba等给出“引用区域”的概念,即包含引用标识符的句子周围的一个连续区域^[10];2010年,Qazvinain等对引文句和上下文进行明确区分,指出引文上下文是一个引用区域内除包含标识符句子(引文据)外的句子集合^[11];2013年,张金松从NLP角度将引文上下文定义为:施引文献为标记处被引文献内容,而在引用标记符号所出现的位置前、后截取的 n 个词^[12];2014年,Parikshit等在总结前人关于引文上下文定义的基础上,提出显式引文上下文和隐式引文上下文的定义,分别对应Nanba和Qazvinain所定义的引文上下文^[13]。由此可见,针对引文上下文的定义有狭义和广义之分,狭义的引文上下文指包含引文标识符的句子;广义的引文上下文包含引文句和在引用区域内引文句前、后句子的集合。

目前国际上关于引文上下文的研究还较少,由于引文上下文的自动识别是引文内容、引文功能、引文情感和引文重要性分析的前提和关键,有关引文上下文的研究主要集中在其自动识别和抽取上。1999年,Nanba等使用引文句中的代词、连接词和人称词等制定引文上下文识别规则和识别引用区域,取得80%的准确率和76%的召回率^[10];Abu-Jbara等采用句法树来改善引文句中含有多个引用的情况^[14];Angrosh等针对文献中相关研究章节的引文上下文,分析该章节的一般引用模式,并将引文上下文的识别转化为分类实验,使用条件随机场进行分类模型训练,最后取得96.51%

的准确率^[15]；2012年，Abu-Jbara等将引文上下文自动识别问题分别转化为单词分类问题、序列标注问题和句子片段分类问题，发现基于句子片段的分类效果最好，取得81.80%的准确率^[16]；2013年，Angrosh使用词汇特征构建CRF模型进行引文上下文识别，并基于此开发引文上下文自动抽取系统CitContExt^[17]；2014年，Sondhi等在构建文献句数-引文数矩阵的基础上，使用隐马尔科夫模型进行引文上下文自动识别^[18]；Athar结合句法特征和词汇特征训练SVM分类器，并证明引文上下文对引文情感和重要性的识别效果可分别提升48%和17%^[19]。

国内关于引文上下文自动识别的研究还处于起步阶段。孙枫军通过识别引文句进行概念抽取研究^[20]；张金松利用基于规则的方法识别引文上下文，并利用引文上下文的语义信息进行文献检索^[12]；雷声伟等梳理引文上下文研究的现状和自动识别研究的不足，归纳出五类特征，分别采用文本分类和序列标注的思想进行引文上下文识别，取得较好效果^[21]。

综上所述，引文上下文虽然提出较早，但相关研究数量还较少；计算机科学、情报学等领域的学者对引文上下文的自动识别研究已取得一定理论成果，但仍存在不足。引文上下文的自动识别主要分为两个方面：一是基于机器学习思想，采用分类、序列标注和条件随机场等模型构建特征工程，进行模型训练和测试，这种方式速度快、自动化程度高，但需要大量人工标注，准确率低；二是基于规则的方式，通过观察和分析引文上下文的行文规律，构建抽取规则，使用正则表达式进行匹配，这种方式准确率高，但需要人工构建抽取规则。为保证引文上下文抽取的准确度，本文采用第二种方式进行引文上下文自动识别。

3 系统思路与构建

科研人员在从事科研创作时，为避免重复性工作，需要广泛地调研和阅读研究领域相关研究成果，并对其核心内容进行归纳、总结和评述。在此情景下，本文假设对于一篇科研文献而言，若有研究人员已对其进行归纳和评述，当这篇科研文献被再次引用时，已有评述可被借鉴使用。因此，本文拟利用引文上下文自动识别技术将某一学科领域科研文献集中相关研究部分的引文上下文识别并抽取，形成文献-引文上下文数据集。在此基础上，从引文上下文的视角实现相关研究的辅

助生成，一方面为科研人员提供基于引文上下文的检索和导航功能；另一方面，通过自动识别得到的相关研究引文上下文组合可快速全面地生成研究初稿，结合用户自定义判断和个性化修改，提升科研效率。此外，通过对引文上下文和文献标题（摘要、全文）进行聚类分析，可进一步帮助科研人员对研究主题相关研究成果的整体态势进行快速把握。对期刊编辑或审稿专家而言，该系统可帮助其检查相关研究中针对某一研究成果的论述是否客观、全面。

3.1 系统思路

为实现基于引文上下文的相关研究辅助生成系统，首先需要解决人工获取研究领域的科研文献集合的问题，识别和抽取出每篇科研文献中的相关研究部分，得到引文上下文集合；在此基础上，对引文上下文进行分词、去停用词等文本预处理，作为检索词从人工收集的领域科研文献集合中获取对应的参考文献及文献题录信息，得到文献集合和对应的文献-引文上下文数据集。其次，将科研人员在书写相关研究时的信息需求划分为根据检索“引文上下文”和检索“相关文献”，对引文上下文和文献题录信息分别建立索引，实现科研文献语句层面的细粒度检索。当用户进行相关文献检索时，输入关键词即可得到相关文献列表；当用户继续点击文献标题时，系统可交互式地返回关于该文献的所有引文上下文集合、摘要及详细的题录信息，利用良好的用户交互体验和对信息需求的细化来优化传统学术数据库的检索结果和使用体验。

为帮助科研工作者更好地完成论文写作，系统需要加入写作模块。用户可自由组合某一研究主题多篇文献的多个引文上下文描述，形成较客观和全面的论文初稿。由于初稿存在内容重复、语法错误等问题，写作模块还应提供相应的编辑模块，使用户可以对初稿的错误进行判断和个性化修改，从而形成具有学者自身科研写作风格的论文终稿。此外，某一研究主题可能含有较多相关研究成果，使检索得到的文献和引文上下文数量超过一定规模，造成浏览困难。为解决这一问题，系统提供相应聚类功能，用户可限定使用文献标题或引文上下文进行聚类。一方面使检索结果分门别类，便于浏览和写作；另一方面，对文献或引文上下文进行聚类，可形成对研究主题的相关研究概览，有助于研究人员对研究主题的整体态势快速地把握。

3.2 系统构建

综合考虑用户的使用情景、需求和现有系统的功

能, 本文实现的相关研究辅助生成系统划分为5个功能模块: 文献检索模块、阅读导航模块、辅助写作模块、聚类分析模块和数据管理模块, 系统功能框架见图1。

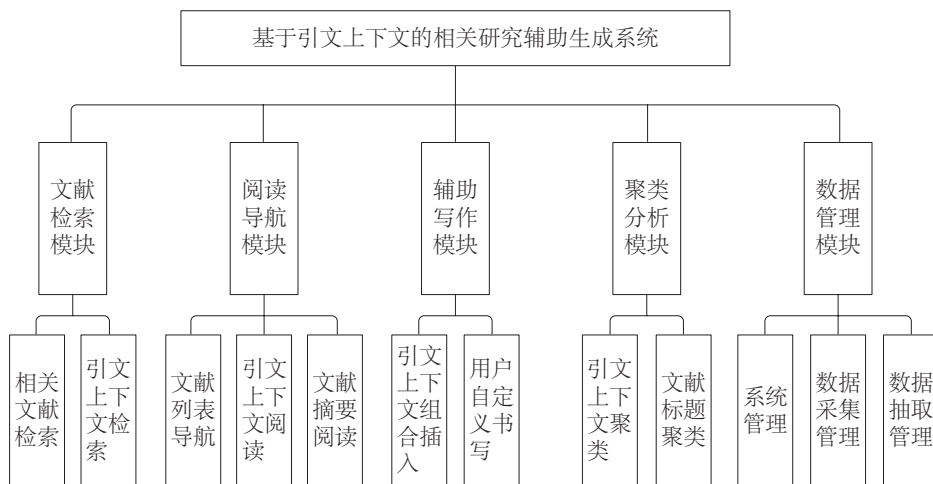


图1 系统功能框架

3.2.1 文献检索模块

本系统的文献检索模块通过分别对领域科研文献的题录信息和引文上下文建立索引, 细化用户信息需求, 以优化传统数据库的检索结果。文献检索模块包含两个子模块(相关文献检索和引文上下文检索), 二者的不同之处在于返回的对象不一样, 前者返回的是与输入的研究主题词或关键词相关的文献信息(包含文献标题、摘要、作者、出版年份等); 后者返回的是相关研究主题的引文上下文列表, 即施引文献中对被引相关文献的描述句。

3.2.2 阅读导航模块

用户在阅读文献时, 通常希望直接阅读主要关注的部分; 科研人员在撰写论文时, 通常希望能直接获取前人撰写的对其所关注文献的描述和评价。阅读导航模块的作用是将这些组织好的信息呈现给用户, 帮助用户快速了解相关研究内容。阅读导航模块包含文献列表导航、引文上下文阅读和文献摘要阅读三个子模块。文献列表导航模块提供文献导航功能, 用户通过点击列表文献, 系统可交互式地返回该文献的摘要和引文上下文。后两个子模块作为容器分别呈现第一个模块的返回值。摘要是科研文献内容的浓缩(包括核心内

容、主要观点和基本情感等), 帮助用户确定文献的利用价值; 多个引文上下文是从更多的角度来阐释和评价被引用文献的主要内容。用户可以通过协调三个子模块, 将线性阅读和非线性阅读结合起来形成交互式阅读, 从而满足其个性化信息需求。

3.2.3 辅助写作模块

本系统直接关注用户撰写相关研究时的实际场景, 并开发了实时辅助写作模块。该模块主要实现引文上下文组合插入和用户自定义书写功能。上下文组合插入功能可帮助用户快速生成某一研究主题的相关研究初稿和对应的参考文献列表; 系统在引文上下文阅读子模块提供选择框, 用户可决定是否包含该引文上下文和该引文上下文在相关研究初稿中出现的相对位置。由于生成的相关研究初稿较粗糙, 用户自定义书写功能允许用户对相关研究初稿进行修改。例如, 修饰润色相关语句、修改行文风格、加入用户对文献的理解等。此外, 用户可随时勾选引文上下文面板中的引文句, 并插入到书写框的光标处。用户通过实现与系统各模块的交互式阅读和写作, 可高效地完成相关研究撰写。此外, 系统还提供自动导出功能, 当用户确认书写完成后, 可直接点击“生成综述”按钮, 系统将自动导出纯文本格式, 方便用户保存和使用。

3.2.4 聚类分析模块

在实际科学研究中,一个研究主题通常涉及多个相关主题的研究内容。例如,图像检索系统的相关研究主题包括用户认知、信息检索和图像语义识别等。基于此,本系统在聚类分析模块提供两种聚类模式,即引文上下文聚类和文献标题聚类。通过聚类分析,用户可快速全面地把握某一研究主题的研究态势。

3.2.5 数据管理模块

数据管理模块主要具备对系统数据进行增添、修改、删除和维护等功能,共包含系统管理、数据采集管理和数据抽取管理三个子模块。系统管理员可通过数

据采集管理模块定期采集数据,对采集的数据进行解析和清洗等操作,通过系统管理模块对已清洗的数据进行修改、维护等,通过抽取管理模块可实现对语句分句、章节抽取规则的修改和增删等。

4 系统实现

本文以计算机领域为例,构建基于引文上下文的相关研究辅助生成系统(Related Works Generation System, RWGS)。RWGS的实现过程分为五个步骤:原始数据采集和数据清洗;引文上下文的识别和抽取,构建文献集、引文上下文集和文献-引文上下文集,并分别建立索引;检索和聚类模块的实现;辅助写作模块实现;系统界面与调试(见图2)。

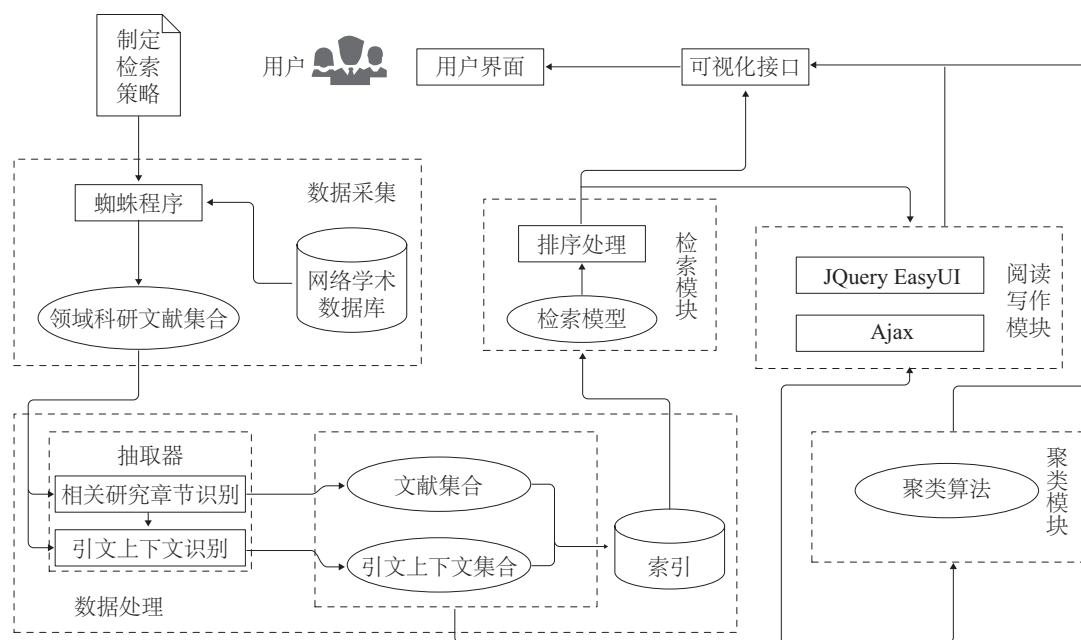


图2 RWGS系统实现过程框架

4.1 原始数据采集和数据清洗

在RWGS的首次数据采集中,本文采用人工收集的方式从Science Direct Onsite数据库中获取计算机领域117本英文期刊在1957—2014年收录的共289 926篇科研文献的全文网页数据。从Science Direct Onsite获取的数据是HTML格式,可避免烦琐的PDF文档解析过程,提高准确率;利用HTML解析器Jsoup对网页无用信息进行过滤,将其处理为便于阅读和爬取的XML

文档。在此基础上,本文结合正则表达式和XML解析器Jdom对文献数据进行信息抽取,获得每篇文献的题录信息(包括标题、摘要、作者、发表时间、关键词等)。同时,将文献的正文数据单独保存,为下一步识别和抽取文献的相关研究章节提供数据基础。此外,由于学术资源的动态性,系统管理员可根据需要对采集的数据进行周期性更新;普通用户可通过数据管理维护模块来手工定制检索策略和采集周期,以构建满足时效性和自身信息需求的领域科研文献数据集。

4.2 文献-引文上下文数据集构建

文献-引文上下文数据集的构建主要依靠抽取器中章节抽取规则和引文上下文抽取规则来实现对正文章节和引文上下文进行识别和抽取。用户可根据不同的科研文献来源定制不同的章节抽取规则和引文上下文抽取规则,还可使用抽取管理模块人工地对抽取规则进行增减、修改和配置。

4.2.1 抽取规则

以*Communications of ACM*刊为例,通过对文献正文的HTML文本进行分析,统计发现大部分研究章节存在于标签“<h1>”和“<p>”下,且章节名称符合一定的规律,如含有“relatedwork”“background”“literature”等词汇。陆伟等在学术文本结构功能识别的系列研究中探讨基于章节标题和段落的学术文本结构识别,并指出若章节标题存在词典 $D=\{\text{background, literatur, framework, previo, relat, measure, method}\}$ 中任一词干,则该章节很大可能是相关研究章节^[22-24]。因此,本文首先从HTML文本中抽取存在于“<h1>”和“<p>”标签下的章节标题集合 $T=\{t_1, t_2, t_3 \cdots t_n\}$,使用波特词干提取法和自然语言处理工具OpenNLP对每个章节标题进行分词和词干提取,得到对应的词干序列 $W=\{w_1, w_2, w_3 \cdots w_m\}$ 。通过观察章节标题中的连接词,本文构建了一个无意义词表 $K=\{k_1, k_2, k_3 \cdots k_x\}$,对每个词干序列进行过滤得到核心词干序列C。若某章节标题的核心词干集中包含词典D中的任一元素,则判定该章节为相关研究章节。

为识别相关研究章节中的引文上下文,需找到含有引用标识符的句子。通过观察发现,引用标识符在正文HTML文本中以“<a>”标签的形式出现。因此,本系统利用OpenNLP实现句子探测,来定向探测相关研究章节包含“<a>”标签的句子,并进行分句处理,从而得到相应的引文上下文集合。其中,为避免一个句子含有多个引用标识符时造成的重复计数,须先对句子进行切分,然后再进行标签探测。

为构建文献-引文上下文数据集,需将文献和其他相关章节中对该文献进行描述的引文上下文对应起来。本系统将引文上下文进行分词、去停用词等文本预处理,然后将结果作为检索词对科研文献集进行查找,得到对应被引用文献。

4.2.2 抽取结果

利用得到的抽取规则,对计算机领域的28万余篇文献进行信息抽取和数据处理,由于部分文献的全文数据缺失,最终得到192 876个不重复的相关研究章节,223 674个不重复的引文上下文句子。为验证抽取的准确性,随机抽取2 000个相关研究章节进行人工标注,得到的准确率为98.52%。由于文献可能多次引用或零引用,最终得到14 501个不重复的文献-引文上下文对。通过对文献集、引文上下文集、文献-引文上下文集分别建立索引,RWGS可将与某一研究主题相关的文献题录信息、引文上下文以及文献-引文上下文信息一同返回给用户。

4.3 检索和聚类模块的实现

在检索模块,RWGS对采集的科研文献进行整合,对元数据(文献的标题、作者、出版年份等)、引文上下文、文献-引文上下文集分别建立索引,并根据需要归并各索引文件。在信息检索过程中,本系统使用传统的信息检索模型和方法来对用户查询进行分词处理、检索和排序,通过可视化结构将结果返回文献列表导航模块。当用户点击文献列表导航的文献时,将激活系统检索模块,RWGS会自动检索其他文献在引用这篇文献时对该文献进行描述的引文上下文和文献摘要。主题聚类模块包括根据标题聚类 and 根据引文上下文聚类,RWGS采用TF-IDF算法计算语句相似度,并通过实现K-means聚类算法来对文献列表和引文上下文列表中的文献标题和引文上下文进行主题聚类。

4.4 辅助写作模块实现

辅助写作模块主要使用Ajax实现用户与系统的交互。本系统设计在检索得到的文献列表点击某篇文献时,将在系统界面中出现该文献的摘要信息以及其他文献的相关引用语句。摘要帮助用户了解全文概况,用户可自行判断此文献是否能满足其信息需求;相关语句的呈现可辅助用户写作。

当引文上下文列表存在满足用户写作需求的条目时,直接勾选该条目,被勾选的语句会出现在写作框,用户即可进入编写模式。当用户写完后,可点击“生成综述”将已编好的内容保存。

4.5 系统界面与运行情况

系统界面的实现采用jQuery EasyUI前端框架,主界面主要包含检索栏、文献列表导航界面、引文句界面、摘要阅读界面、写作面板、参考文献界面。用户在检索框输入检索词进行相关文献搜索时,文献列表导航界面会返回相关文献的标题、作者和发表年份等信息,供用户浏览和选择。点击作者,结果按照作者名字的首字母升序排列;点击标题,结果按标题第一个单词的首字母升序排列;点击年份,结果按发表年份降序排列。当用户点击某篇感兴趣的文献标题时,引文句界面将返回其引文上下文。此外,系统还提供“语句分组”和“文献分组”功能,分别用于激发基于引文上下文的主题聚类 and 基于文献标题的主题聚类,来应对文献和引文上下文较多的情况,帮助用户全面、高效地完成论文写作。

5 结语

本文从引文上下文的角度出发,设计基于引文上下文的相关研究辅助生成系统,并在计算机领域对该系统进行实现。虽然系统在一定程度上改善了已有研究成果,但仍然存在很多不足,需要进一步完善。本文将在后续研究中对原始数据集进行扩充,使其兼容更多的数据源,构建更大的文献-引文上下文数据集,以提高系统结果的全面性和可靠性。针对主题聚类模块,目前实现了基于引文句和文献标题的聚类,下一步将考虑基于文本标题和摘要的聚类。此外,本系统在引文上下文的抽取中,为突出研究重点和保证抽取的准确性,采用基于规则模板的引文句抽取,下一步应尝试基于机器学习和深度学习的引文上下文句子集合的抽取,从而增加系统的适应性和灵活性。

参考文献

- [1] DEMCHENKO Y,ZHAO Z,GROSSO P,et al.Addressing big data challenges for scientific data infrastructure[C]//IEEE,International Conference on Cloud Computing Technology and Science.Taipei: IEEE Computer Society,2012:614-617.
- [2] 陈超美,陈悦,侯剑华,等.CiteSpace II:科学文献中新趋势与新动态的识别与可视化[J].情报学报,2009,28(3):401-421.
- [3] HEIMERL F,HAN Q,KOCH S,et al.CiteRivers:visual analytics

- of citation patterns[J].IEEE Transactions on Visualization & Computer Graphics,2015,22(1):1.
- [4] CHEN M H,HUANG S T,HSIEH H T,et al.Flow:a first-language-oriented writing assistant system[J].ACL System Demonstrations, 2012,24(3):157-162.
- [5] 杨秉哲.WriteAhead:以学术论文写作为目的之摘要写作辅助系统[D].新竹:台湾清华大学,2009.
- [6] VAN ECK N J,WALTMAN L.Vosviewer:a computer program for bibliometric mapping[J].Social Science Electronic Publishing,2009, 84(2):523-538.
- [7] WANG X,CHENG Q K,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.
- [8] 孔行.基于主题推荐的辅助写作系统[D].哈尔滨:哈尔滨工业大学,2015.
- [9] CHUBIND E,MOITRA S D.Content analysis of references:adjunct or alternative to citation counting?[J].Social Studies of Science,1975,5(4): 423-441.
- [10] NANBA H,OKUMURA M.Towards multi-paper summarization using reference information[J].Ipsj Sig Notes,1999,98(82):79-86.
- [11] QAZVINIAN V,RADEV D R.Identifying Non-explicit Citing Sentences for Citation-based Summarization[C]//ACL 2010,Proceedings of the Meeting of the Association for Computational Linguistics,July 11-16, 2010,Uppsala,Sweden.DBLP,2010:555-564.
- [12] 张金松.基于引文上下文分析的文献检索技术研究[D].大连:大连海事大学,2013.
- [13] SONDHI P,ZHAI C X.A constrained hidden markov model approach for non-explicit citation context extraction[M]//Proceedings of the 2014 SIAM International Conference on Data Mining.Pennsylvania: Society for Industrial and Applied Mathematics,2014:361-369.
- [14] ABU-JBARA A,RADEV D.Coherent citation-based summarization of scientific papers[C]//Meeting of the Association for Computational Linguistics:Human Language Technologies.Portland:DBLP, 2011:500-509.
- [15] ANGROSHM A,CRANFIELD S,STABGER N.Context identification of sentences in related work sections using a conditional random field:towards intelligent digital libraries[C]//Proceedings of the 10th annual joint conference on Digital libraries,Gold Coast: ACM,2010:293-302.
- [16] ABU-JBARA A,RADEV D.Reference scope identification in citing sentences[C]//Proceedings of the 2012 Conference of the North American Chapter of ACM.Montreal:ACM,2012:80-90.
- [17] ANGROSHM A,CRANFIELD S,STANGER N.Context

- identification of sentences in research articles:towards developing intelligent tools for the research community[J].Natural Language Engineering,2013,19(4):481-515.
- [18] SONDHI P,ZHAI C X.A constrained hidden markov model approach for non-explicit citation context extraction[C]// Proceedings of the 2014 SIAM International Conference on Data Mining. Philadelphia, Pennsylvania, 2014:102-108.
- [19] ATHAR A.Sentimental analysis of scientific citations[EB/OL]. [2017-07-01].<http://www.cl.cam.ac.uk/techreports/UCAM-CL-TR-856.pdf>.
- [20] 孙枫军.引文上下文中的概念抽取[D].北京:中国科学技术信息研究所,2012.
- [21] 雷声伟,陈海华,黄永,等.学术文献引文上下文自动识别研究[J].图书情报工作,2016(17):78-87.
- [22] 陆伟,黄永,程齐凯,等.学术文本的结构功能识别功能框架及基于章节标题的识别[J].情报学报,2014(9):979-985.
- [23] 黄永,陆伟,程齐凯,等.学术文本的结构功能识别——基于段落的识别[J].情报学报,2016,35(5):530-538.
- [24] 黄永,陆伟,程齐凯,等.学术文本的结构功能识别——在学术搜索中的应用[J].情报学报,2016,35(4):425-431.

作者简介

王鑫,男,1996年生,硕士研究生,研究方向:信息检索。

程齐凯,男,1989年生,博士,讲师,研究方向:自然语言处理、文本挖掘、信息检索, E-mail: cehngqikai0806@163.com。

李信,男,1991年生,博士研究生,研究方向:大数据分析、语义计量、医学知识发现, E-mail: lucian@whu.edu.cn。

陆伟,男,1974年生,教授,博士生导师,研究方向:信息检索、文本挖掘和知识发现, E-mail: weilu@whu.edu.cn。

Design and Implementation of Related Works Generation System Based on Citation Context

WANG Xin^{1,2}, CHENG QiKai^{1,2}, LI Xin^{1,2}, LU Wei^{1,2}

(1. School of Information Management, Wuhan University, Wuhan 430072, China;

2. Information Retrieval and Knowledge Mining Laboratory, Wuhan University, Wuhan 430072, China)

Abstract: This article takes specific information needs when a scholar is writing related works or a edit reviews a paper to design and implement a related works assistant system based on citation context. Firstly, we expound the definition of citation context in academic texts and the progress of its recognition, then discuss the feasibility of implementation a system like this and design the thought and functional modules of it. Finally, the full text page data of 289 926 scientific literatures included in the Science Direct database from 1957 to 2014 were used as data sources to realize a related research assistant generation system RWGS based on citation context. The result shows that RWGS can meet the needs of scholars in the preparation of the related works chapter or journal editor in the review process with more detailed information needs, which have a certain optimization effect, while the literature search the traditional academic database search results.

Keywords: Related Works; Assistant Generation System; Citation Context; Computer Science

(收稿日期: 2017-08-09)