

基于 XML 全文数据引文分析系统的设计与实现^{*}

胡志刚^{1,2} 陈超美^{1,3} 刘则渊^{1,2} 侯海燕^{1,2}

¹(大连理工大学-德雷塞尔大学知识可视化与科学发现联合研究所 大连 116024)

²(大连理工大学 WISE 实验室 大连 116024)

³(德雷塞尔大学信息科技学院 费城 19104)

【摘要】构建一种基于 XML 全文的引文分析系统,可以在施引文献的全文中识别和提取出引用的位置、引用的语境等信息,并将提取到的引用信息存储到一个关系数据库中,以供被引检索和引文分析。结合纳米管的案例,对一篇引文进行被引检索和检索结果分析实验,验证本系统的可行性和执行效果,同时也显示该系统可以为传统的引文分析提供一种微观视角和基于全文数据的引文分析和查询工具。

【关键词】引文分析 XML 全文 引用位置

【分类号】G354.4

Design and Implementation of the Citation Analysis System Based on XML Full-text Articles

Hu Zhigang^{1,2} Chen Chaomei^{1,3} Liu Zeyuan^{1,2} Hou Haiyan^{1,2}

¹(Joint-Institute for the Study of Knowledge Visualization and Science Discovery,

Dalian University of Technology(China)-Drexel University(USA), Dalian 116024, China)

²(WISELab, Dalian University of Technology, Dalian 116024, China)

³(College of Information Science and Technology, Drexel University, Philadelphia 19104, USA)

【Abstract】In this paper, the authors design a novel citation analysis system based on the full-text literature, which can be used to identify citation location, cited references, and citation context as well as other citation information in the full text. The extracted information is imported into a relational database for the following citation retrieval and citation analysis. A case study is conducted using the field of Nano-tube, cited retrieval is tested firstly and then an analysis based on the search results is conducted. This experiment shows the function and the effect of the system and proves its performance on citation analysis in micro-level and based on full-text data.

【Keywords】Citation analysis XML full-text Citation location

1 引言

随着互联网技术的发展,结构化的文献全文信息越来越容易获得。作为一种重要的信息资源结构化描述方式,XML 全文格式已经成为各文献数据库的一种重要的全文显示方式,Springer、Elsevier 和 Wiley 都提供或部分

收稿日期:2012-11-09

收修改稿日期:2012-11-13

* 本文系国家社会科学基金重大项目“高科技伦理问题的善治:理论与战略框架”(项目编号:12ZD117)的研究成果之一。

提供 XML 格式的全文阅读或下载。比如, Elsevier 运行的 ConSyn 数据库提供了 XML 格式全文的批量下载, 知名开放获取 (Open Access) 出版物 PLOS ONE, 在传统的 PDF 下载之外, 也提供了 XML 格式的全文下载。

全文数据的出现也给传统的引文分析带了新的条件和可能。传统的引文分析, 受题录式引文数据的限制, 只能研究文献之间的引用或被引关系, 无法识别被引文献在施引文献中的引用位置、引用语境等信息。而本文在 XML 全文引文数据的基础上, 利用 PHP + MySQL 语言, 设计和实现了一种基于服务器和浏览器方式的引文分析系统, 通过对施引文献的 XML 格式全文的解析, 可以识别和分析施引文献中的引用位置、引用语境等信息, 从而丰富了传统引文分析的功能和用法。

2 研究背景

文献计量学正沿着从题录分析到引文分析, 再到全文分析的趋势向前发展。题录分析是利用文献数据库所提供的文献题录信息对文献进行分析, 如作者分析、词频分析等。引文分析是利用引文数据库, 通过对引文的分析 (如被引次数的统计), 来挖掘某一领域的高被引文献及其学术影响。然而, 传统的引文分析由于数据本身的限制, 只能给出一篇文献被引用的次数, 并不能反映文献被引用的具体语境, 如引文在施引文献中的什么位置被引用, 在一篇文献被引用了几次, 每次引用的具体语境是什么。Garfield、Moed 等引文分析的早期开创者也指出, 单纯考虑数量的引文分析方法具有局限性^[1-4], 因为它没有考虑到引用行为和动机的复杂性和多样性^[5,6]。

要解决这些问题, 就需要施引文献的全文数据和施引文献的全文分析^[7-9]。比如, 就引用位置而言, Cano^[10]发现引用的位置因素与引用的行为有关; Frøsig^[11]也认为引用位置所在的章节是对引用行为进行分类的一个重要参考; Herlach^[12]认为单纯的被引次数指标过于简单, 还应该考虑到一篇引文在文献中被引用或提及的次数; Teufel 等^[13]则指出对引用的行为进行识别需要找到引文的引用语境, 通过引用语境的分类对引用的行为和动机进行分类^[14-16], 同时改善文献被引检索的性能和精度^[17-20]。

之前由于全文数据的不易获取和不易解析, 基于

全文的引文分析, 尤其是大样本的实证研究比较少见。在互联网时代, 科学文献全文数据尤其是结构化的全文数据 XML, 变得越来越易得。相对于 PDF 格式的全文, XML 格式的文献全文具有以下特点^[21,22]:

(1) 结构化: XML 是一种结构化的标识语言, 更易于标识文章的题录信息和引文的位置和上下文信息;

(2) 通用性: XML 是一种由浏览器支持的通用格式, 不受软件和平台的限制, 并且可以自定义各种丰富的显示样式;

(3) 交互性: XML 中可以包含丰富的超链接, 以方便在文章或数据库中跳转, 大大提高了文章的交互性和数据库的连通性。

Elsevier 的 XML 全文数据格式是影响较广的一个全文元数据格式, 开放获取期刊 PLOS ONE 就采取了 Elsevier 的 XML 数据格式作为数据的存储和中转中介。Elsevier 的 XML 格式数据的文档类型定义 (Document Type Definition, DTD) 和 XML 架构 (Schema) 的具体描述可以从 Elsevier 的官方网站^①上获取, Elsevier ConSyn 数据库提供 XML 全文数据的检索和下载。

在 Elsevier 的 XML 全文数据格式中, 对正文中出现的引用使用超链接和标识符的方式进行了标引, 可以方便地在施引文献的全文中识别出引用的位置等信息, 这为本文在微观层次上研究引文提供了条件。正是基于 Elsevier 的 XML 全文数据, 利用 PHP 语言在处理 XML 数据上的优势, 本文构建并开发了一种引文分析系统, 以实现基于全文的引文分析的大样本进行实证研究。

3 系统设计和实现

本系统主要由两个功能模块构成, 如图 1 所示:

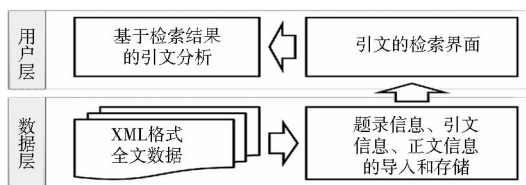


图 1 基于全文的引文分析系统的实现流程和功能模块

① http://www.elsevier.com/wps/find/authorsview_authors/dtds.htm

(1)数据层:在服务器端,首先进行 XML 全文数据的解析和数据库存储;

(2)用户层:在浏览器端,设计了引文的检索、筛选和结果界面,同时为了进一步的计量分析,还提供对检索结果的导出(以表格的形式)。

3.1 数据的解析和存储

一个标准的 Elsevier XML 全文数据由题录信息、正文信息和引文信息三部分组成。其中,题录信息和引文信息的导入比较简单,可以利用 PHP 中的 SimpleXML 函数直接进行解析和提取;而正文信息的解析需要经过对正文全文的遍历,较为复杂,是本系统的难点和重点,如图 2 所示:

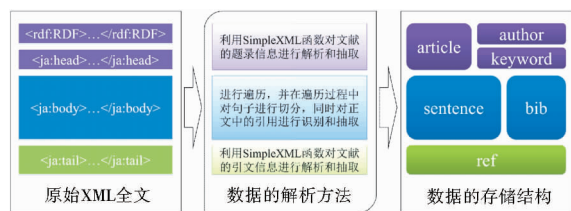


图2 XML全文数据的解析和存储流程

(1)题录信息的解析和存储

题录信息的提取较为简单,可以利用 PHP 中的 SimpleXML 函数,将 XML 字符串进行解析,并载入到对象变量 \$object 中,题录信息主要在 \$object→rdf_RDF→rdf_Description 或者 \$object→\$ja_article→\$ja_head 两个对象中,其中包括了该文献的标题(dc_title)、作者(dc_creator)、期刊(prism_publicationName)、年份(prism_coverDate)、期卷(prism_volume)、起止页码(prism_startingPage, prismendingPage)、关键词(dc_subject)、摘要(ce_abstractsec)等。

需要注意的是,由于 Elsevier XML 格式的 XML 标签中含有冒号,如 <rdf:RDF>,会导致 SimpleXML 函数无法正常运行,因此需要先将标签中的冒号替换为“_”或其他字符。

在提取题录数据之后,需要将它们存储在一个 MySQL 数据库中。根据上述各题录信息之间的关系,本文设计了三个数据表进行存储,分别是 article、author、keyword。其中,除作者和关键词外的各信息存储在 article 数据表中;作者和关键词信息因为与文章之间存在多对多的关系,因此分别存储在 author 和 keyword 表中,并以文章序号与 article 建立索引关系。

(2)引文信息的解析和存储

同题录信息的提取类似,引文信息主要存在 ja_tail 对象下面的 ce_bibliography 中,包括每个引文的标题(sb_title)、作者(sb_authors)、期刊(sb_series)、年份(sb_date)、期卷(sb_volumenr)、起止页码(sb_pages)等。如果引文的类型是图书(sb_book)、编集(sb_editedbook)或其他类型(ce_otherref),引文的信息略有不同。提取的引文信息会被存储在 ref 数据表中。

(3)正文信息的提取和存储

正文信息的解析最为复杂,由于在 Elsevier XML 格式的全文中,正文的基本单元是段落(ce_para),而不是句子,因此需要通过遍历的方法将段落切分成句子,同时在遍历过程中标识出文中可能存在的引用,如图 3 所示:

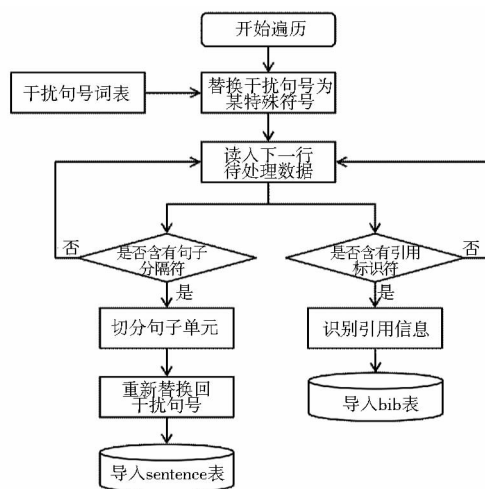


图3 正文的遍历流程与正文信息的提取

在对正文的遍历中,使用句号(.)和问号(?)作为切分句子的标志。感叹号(!)虽然也可以用来切分句子,但因为在学术论文中感叹号的使用非常罕见,为了保证程序的运行效率,不将其作为句子的切分标志。另外,由于句号(.)除作为句子结束符外,还可能出现在人名(如“Iijima S.”)、数字(如“0.123”)或其他缩写中(如“etc.”、“e.g.”、“Fig. 1”)中,对于这类情况,采取词表替换(主要针对缩写中的句号)和正则表达式替换(主要针对人名和数字中的句号)相结合的方法,将干扰句号首先替换为其他特殊符号,切分之后再恢复。切分得到的句子,会依次存储在 sentence 数据表中,每个句子作为数据表中的一条记录,存储字

段主要包括句子的长度和位置,包括所在的小节(section)、段落(paragraph)等。

另外,在正文的遍历过程中还要完成对引用信息的提取。当遇到引用的标识“<ce:cross-ref>”时,系统将其视为一条引用信息,并记下当前的位置作为该次引用的位置(包括所在的小节、段落、句子以及单词数),同时,根据引用标识中的属性信息,如“refid = bib2 bib3”,找到该次引用所引用的引文。引用信息被存储在 bib 数据表中,每条引用信息对应数据表中的一条记录。

以上 6 个数据表之间存在着如下的一对多或者多对多的关系,如图 4 所示:

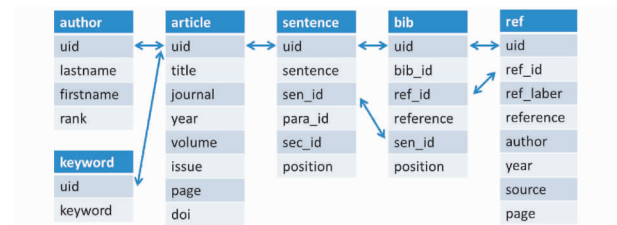


图 4 基于 XML 全文的引文分析系统底层数据库表结构和关系

author 和 article 之间存在多对多的关系,一篇文章可以有多个作者,而一个作者也可以发表多篇文章;keyword 和 article 之间存在类似的关系;article 和 sentence 之间的关系是一对多的,一篇文章对应多个句子,而一个句子只可能存在于一篇文章中;sentence 和 bib 之间的关系也是一对多的,一个句子中可以有一个或多个引用,而一个引用只能存在于一个句子中;bib 和 ref 之间存在多对多的关系,一个引用位置可能引用多篇引文,而一篇引文可以在多个引用位置被引用。

3.2 引文的筛选和检索

在完成全文数据的导入和存储之后,可以通过数据库的查询功能进行引文的检索,如图 5 所示。

本系统参考 ISI Web of Science 的浏览器端用户界面和检索流程,设计两步式的检索策略,即首先根据用户提交的检索项,查询并返回所有可能的引文供用户进行筛选,然后经由用户勾选提交后系统再进一步查询被引文献的施引信息。与 ISI Web of Science 不同的是,该系统返回的结果不再是施引文献,而是一条条的引用信息,即引文在施引文献中的具体引文位置、引用语境等。

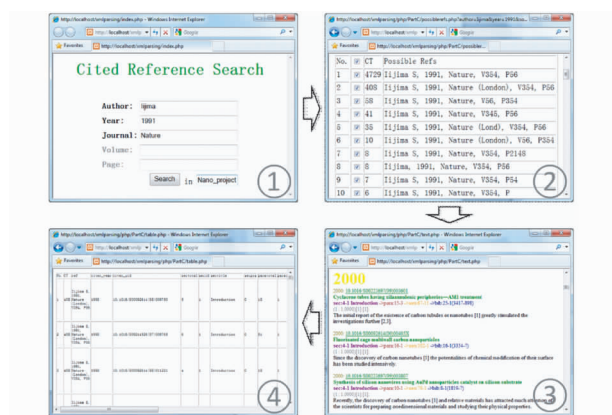


图 5 基于 XML 全文的引文分析系统的被引检索界面

(1) 引文的查询和筛选

由于引文的格式通常比较杂乱,本系统设计了引文筛选的中间步骤,中间筛选过程可以大大提高被引检索的查全率和查准率。在这一步中,用户首先填写想要检索的引文的作者、年份和期刊信息(出于其他检索目的,也可以只填写其中的一项或两项),客户端将表单提交给服务器端(如图 5 - ①所示),服务器端根据提交的检索项生成 SQL 语句,在 ref 数据表中查询所有可能的引文。所生成的 SQL 语句是:select reference from ref where author like 'xxxx%' [and year = xxxx [and source like 'xxxx%']]。服务器将利用这一 SQL 查询得到的记录列表,按照被引次数的高低进行排序后,返回给用户(如图 5 - ②所示)。用户根据服务器返回的可能的引文列表,判断它们是否为所要查找的引文并进行勾选,然后再次提交服务器端进行第二步检索。

(2) 引用信息的检索

在这一步中,服务器根据用户提交的引文列表进行被引检索。该步检索需要在 bib、sentence 和 article 三个数据表中进行,首先通过在 bib 数据表的查询得到引文的施引文献编号及其在施引文献中的具体位置(如所在的句子编号),所用 SQL 语句为 select uid, sen_id from bib where ref_id in ('refid1', 'refid2', ...);然后在 article 数据表根据施引文献的编号给出该施引文献的题目和 DOI 等信息,所用 SQL 语句为 select * from article where uid = uid_value;同时在 sentence 数据表中根据句子编号给出该句子的内容,即引用的语境信息,所用 SQL 语句为 select * from sentence where uid

= uid_value and sen_id = senid_value。

检索得到结果如图 5-③所示,每条结果为一引用信息,包括引用所在的施引文献、在施引文献中的位置和具体语境。为了方便进行其他统计分析,系统还提供对检索结果的表格显示和导出,如图 5-④所示。

4 实验——以纳米管为例

本文以纳米管(Nano Tube)领域的 48 665 篇全文数据为例,对基于 XML 全文的引文检索系统进行实验。这些数据于 2012 年 8 月 9 日下载自 Elsevier 的 ConSyn 全文数据库,数据总大小(zip 压缩)为 882MB。首先,将这些全文导入 MySQL 数据库中,大约需要 5 小时(与服务器端配置的高低有关)。然后,选取日本电镜专家 Iijima 在 1991 年发表在 Nature 杂志上的一关于碳纳米管的高被引论文进行引文检索。在对应的文本框里分别输入“Iijima”、“1991”和“Nature”进行检索。

提交检索,服务器端返回所有符合检索条件的引文,并按照被引次数的从高到低进行排列(如图 5-②所示)。显然,排在前 10 的这些引文所指的都是所要的引文,只是格式上略有差异,因此将这些引文全部进行勾选,然后再次提交进行第二步查询。

在第二步中,系统根据用户提交的引文列表,查询并返回它们的施引信息,共 3 616 条。图 6 列出了返回的其中 4 条引用信息的具体组成和形式,共分 5 行,分别是:施引文献的年份和 DOI 链接、施引文献的标题、该次引用在施引文献中的引用位置、引文在施引文献中的被引强度(即在该施引文献被引用的次数)以及该次引用的具体语境。

在图 6 的引用信息中显示了每次引用的具体位置,以章节序号、段落序号、句子序号和单词序号来测度。例如,在图 6 中第 2 条引用信息显示,该次引用位于施引文献的第 1 节(共 4 节)Introduction,第 2 段(共 11 段),第 12 句(共 78 句)和第 310 个单词(全文共 3 361 个单词)处,是全文 25 个引用位置中的第 2 个。

利用检索结果的导出功能,将检索到的 3 616 条引用位置导出到 Excel 中,并对引用位置进行统计。

通过对 Iijima 的引用位置的统计发现,高达 98% 的引用出现在 Introduction 中,大约 2% 的引用出现在 Results and Discussion 中,几乎没有对 Iijima 的引用是出现在 Experimental 和 Conclusions 中,如图 7-①所

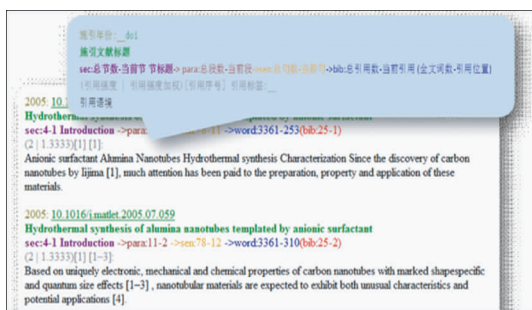


图 6 Iijima 一文的被引检索结果及含义

示。这既反映了一般引用的位置特点,也反映了 Iijima 自身的被引特点,即 Iijima 一文被引的目的是用来进行背景的表述(在 Introduction 的引用),而不是方法的阐述(在 Experimental 中的引用)。

图 7 中还显示了按照段落、句子、单词划分的引用位置的分布。比如,将全文按照单词分为 10 等份,那么,2 133 次即 59% 对 Iijima 的引用,出现在全文的第 1 等份,即全文的前 10% 的位置;25% 的引用出现在第 2 等份,即 10% - 20% 之间的位置;8% 的引用位于第 3 等份;其余位置的引用只占 8%。这说明,对 Iijima 的引用在引用位置上具有极大的不均衡性,大部分的引用位于全文的开始部分。

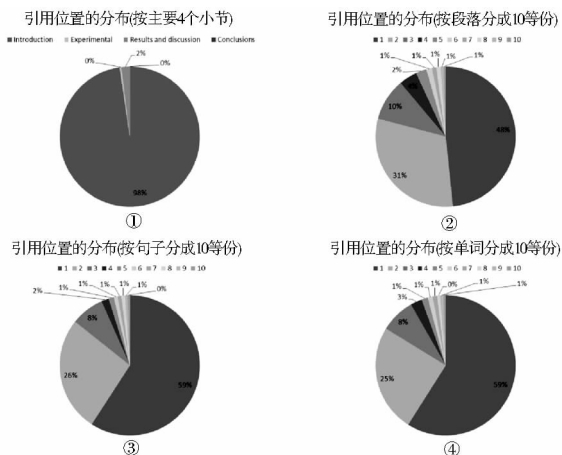


图 7 Iijima 一文被引的引用位置分布

5 结 语

传统的引文分析,通常由于数据所限而局限于宏观层面,对于微观范围内的各类特性缺少统一的分析框架。针对越来越普遍的结构化全文数据,引文分析需要向更为精细的方向发展。本文构建了一种基于

XML 全文的引文分析系统,可以在完成引用信息的解析、提取和存储的基础上,实现被引筛选和检索,得到被引的语境信息,并可以根据这些引用信息,对引用位置等进行计量统计。

从 Iijima 关于碳纳米管一文的被引检索来看,对该文的引用大部分位于 Introduction 一节中,具体位置一般位于全文开始的前 10% 部分。这一实验结果,验证了本系统的可行性和执行效果,同时也显示了该系统可以为传统的引文分析提供一种微观视角的引文检索和分析工具。

(致谢:感谢 Elsevier ConSyn 提供了 XML 格式全文的海量数据。本研究部分工作是在美国德雷塞尔大学联合培养期间完成的,感谢德雷塞尔大学信息科技学院和国家留学基金委为此提供了科研条件和资金资助。)

参考文献:

- [1] Chubin D, Garfield E. Is Citation Analysis a Legitimate Evaluation Tool[J]. *Scientometrics*, 1980, 2(1):91-94.
- [2] Moed H F. Citation Analysis in Research Evaluation [M]. Netherlands: Springer, 2005.
- [3] Bornmann L, Daniel H. What do Citation Counts Measure? A Review of Studies on Citing Behavior[J]. *Journal of Documentation*, 2008, 64(1):45-80.
- [4] MacRoberts M H, MacRoberts B R. Problems of Citation Analysis: A Critical Review[J]. *Journal of the American Society for Information Science and Technology*, 1989, 40(5):342-349.
- [5] Liu M X. Progress in Documentation the Complexities of Citation Practice: A Review of Citation Studies[J]. *Journal of Documentation*, 1993, 49(4):370-408.
- [6] Case D O, Higgins G M. How Can We Investigate Citation Behavior? A Study of Reasons for Citing Literature in Communication [J]. *Journal of the American Society for Information Science*, 2000, 51(7):635-645.
- [7] White H D. Citation Analysis and Discourse Analysis Revisited [J]. *Applied Linguistics*, 2004, 25(1):89-116.
- [8] Smith L C. Citation Analysis[J]. *Library Trends*, 1981, 30(1):83-106.
- [9] Yu H, Agarwal S, Frid N. Investigating and Annotating the Role of Citation in Biomedical Full-text Articles[C]. In: *Proceedings of the 2009 IEEE International Conference on Bioinformatics and Biomedicine Workshop (BIBMW'09)*. Washington, DC: IEEE Computer Society, 2009:308-313.
- [10] Cano V. Citation Behavior: Classification, Utility, and Location [J]. *Journal of the American Society for Information Science*, 1989, 40(4):284-290.
- [11] Frøsig R E. Citation Classification Based on Genre: The Significance of the Textual Location of Citations[D]. Copenhagen: Royal School of Library and Information Science, 2011.
- [12] Herlach G. Can Retrieval of Information from Citation Indexes be Simplified? Multiple Mention of a Reference as a Characteristic of the Link Between Cited and Citing Article[J]. *Journal of the American Society for Information Science*, 1978, 29(6):308-310.
- [13] Teufel S, Siddharthan A, Tidhar D. Automatic Classification of Citation Function[C]. In: *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP'06)*. Stroudsburg, PA: Association for Computational Linguistics, 2006:103-110.
- [14] Garzone M, Mercer R E. Towards an Automated Citation Classifier [C]. In: *Proceedings of the 13th Biennial Conference of the Canadian Society on Computational Studies of Intelligence: Advances in Artificial Intelligence*. London, UK: Springer-Verlag, 2000:337-346.
- [15] Pham S B, Hoffmann A. A New Approach for Scientific Citation Classification Using Cue Phrases[C]. In: *Proceedings of the 16th Australian Conference on Artificial Intelligence*. Berlin, Heidelberg: Springer-Verlag, 2003:759-771.
- [16] Radoulov R. Exploring Automatic Citation Classification[D]. Waterloo: University of Waterloo, 2008.
- [17] Ritchie A, Teufel S, Robertson S. How to Find Better Index Terms Through Citations[C]. In: *Proceedings of the Workshop on How Can Computational Linguistics Improve Information Retrieval (CLI-IR'06)*. Stroudsburg, PA: Association for Computational Linguistics, 2006:25-32.
- [18] Aljaber B, Martinez D, Stokes N, et al. Improving MeSH Classification of Biomedical Articles Using Citation Contexts [J]. *Journal of Biomedical Informatics*, 2011, 44(5):881-896.
- [19] Giles C L, Bollacker K D, Lawrence S. CiteSeer: An Automatic Citation Indexing System[C]. In: *Proceedings of the 3rd ACM Conference on Digital Libraries (DL'08)*. New York: ACM, 1998:89-98.
- [20] Ritchie A, Robertson S, Teufel S. Comparing Citation Contexts for Information Retrieval[C]. In: *Proceedings of the 17th ACM Conference on Information and Knowledge Mining (CIKM'08)*. New York: ACM, 2008:213-222.
- [21] Amer-Yahia S, Lakshmanan L V S, Pandit S. FlexPath: Flexible Structure and Full-text Querying for XML[C]. In: *Proceedings of the 2004 ACM SIGMOD International Conference on Management of Data (SIGMOD'04)*. New York: ACM, 2004:83-94.
- [22] Amer-Yahia S, Shanmugasundaram J. XML Full-text Search: Challenges and Opportunities[C]. In: *Proceedings of the 31st International Conference on Very Large Data Bases (VLDB'05)*. VLDB Endowment, 2005:1368.

(作者 E-mail: huzhigang.cn@gmail.com)