中国科学技术信息研究所

硕士学位论文

引文上下文中的概念抽取

Concept Extraction in Citation Context

作者　孙枫军

导师　朱礼军

中国科学技术信息研究所

论文提交日期(2012年10月)

中图分类号 TP391　　学校代码 80901
UDC　　密　　级 公开

中国科学技术信息研究所

硕士学位论文

引文上下文中的概念抽取

Concept Extraction in Citation Context

作者姓名 孙枫军　　学　　号 809011013
导师姓名 朱礼军　　职　　称 副研究员
学位类别 管理学　　学位级别 硕　士
学科专业 情报学　　研究方向 知识工程

中国科学技术信息研究所

论文提交日期(2012年10月)

独 创 性 声 明

本人郑重声明：所呈交的学位论文，是本人在导师指导下，独立进行研究工作所取得的成果。尽我所知，论文中除已经加以标注和致谢的地方外，不包含任何他人享有著作权的内容。对本论文所涉及的研究工作做出贡献的其他个人和集体，均已在文中明确说明并表示了谢意。

研究生签名：孙枫军　　时间：2012 年 10 月 17 日

关于论文使用授权的说明

本人完全了解中国科学技术信息研究所有关保留、使用学位论文的规定，即：所里有权保留送交论文的打印稿和电子稿，允许论文被查阅和借阅，可以采用影印、缩印或扫描等复制手段保存、汇编学位论文。同意中国科学技术信息研究所用不同方式在不同媒体上发表、公布论文的全部或部分内容。保密的论文在解密后遵守此规定。

研究生签名：孙枫军　　时间：2012 年 10 月 17 日

导师签名：　　时间：2012 年 10 月 17 日

致 谢

两年半的时光飞逝而去，无限感慨，无限怀念。我认为，这两年半的研究生学习生涯是我人生目前为止最为幸福、最为重要的时光。在这段时间里，我找回了自信，也找到了人生的方向，我想今后的人生道路上我也将不再彷徨。更加重要的是，我也认识了许许多多和蔼的老师和可爱的同学，我想，这也将是我的一笔最为宝贵的财富。

首先，我要感谢导师朱礼军老师，朱老师是我的导师，但同时又像是一位兄长。您不仅在学习上关心我们，在生活上也处处替我们着想。毕业论文的完成不仅仅只是我们的付出，更是您在两年半时间里细心耕耘的结果。我在此衷心地向您表示感谢。

我要感谢技术支持中心的刘耀、徐硕、张云良、李颖、乔晓东、闫莹莹、于薇等老师。他们都在我需要的时候给予了我无私的帮助。

我要感谢研究生部的王桂凤老师、赵志耘所长、罗勇主任、张泽玉主任、赵琳老师、刘敏老师、郝文英老师，你们的付出才使得我们能够拥有良好的学习和生活条件。

我还要衷心地感谢武夷山所长给予我的一次又一次的帮助。

我要感谢身边的同学，我不会忘记大家一起生活和学习的日子，我会珍惜这宝贵的同窗之情，永远铭记在心。

最后感谢我的家人，在我求学生涯中一直默默地关心和支持，使我能够顺利的完成学业，感谢你们的包容和帮助，你们无私的支持是我不断进取的永恒动力。

引文上下文中的概念抽取

摘　要

科技文献区别于其它同样以自然语言形式存在的文档的重要特征在于科技文献包含参考文献，引文符号前后一个较小区域内的文本段被称为引文上下文。在较长的一段时间里，引文上下文的文本处理都没有得到足够的重视。然而，随着计算机技术的发展、科技文献可读文本化的实现以及科技文献开放获取运动的发展，对引文上下文的大规模计算机化处理已经成为了可能，引文上下文的研究工作也因此将迎来快速发展的阶段。

在对引文上下文的概念抽取的研究现状加以阐述的基础上，针对引文上下文的概念抽取难以实现自动化的问题，本文提出了引文上下文中概念抽取的方法，设计了引文上下文的概念抽取的系统，系统能够在限定条件下解决引文上下文中概念抽取自动化的问题，可以覆盖全部的参考文献和施引文献。而后，选取两年共计 455 篇某一期刊的文章作为实验数据，进行针对系统的实验，抽取了期刊文章对应参考文献的引文上下文当中以名词性短语形式存在的概念。结果表明，该系统能够达到接近自动化抽取概念的程度，并且可以覆盖研究范围内的全部参考文献和施引文献。

图 2 幅，表 17 个，英文参考文献 41 篇。

关键词：引文上下文；概念抽取；引文数据库；引文索引

分类号：TP391

Concept Extraction in Citation Context

Abstract

Citation index is widely used in scientific literature system and scientometrics, but a problem ignored by many is that citation index lacks the support of content in both information retrieval systems and scientometrics, in other words, the current citation index connects the scientific papers but not the information inside the paper. For information retrieval, compared with subject index, citation index can in some sense avoid the semantic difficulties, but it is difficult to say how the citing or cited documents are related with the seed document and which one of the citing or cited documents satisfy the user's information need. For scientometrics, the basic unit commonly used today is still scientific document, but what really matters in the scientific development is the content in the paper, namely, the important concept, the widely used database, tool or method. As the development of computer technology and open access movement, the time is ripe to extract the important concepts in the citation context for both information retrieval and scientometrics.

Based on the review of current research status and key technology, this study proposes a method of automatic concept extraction in the citation context. To overcome the difficulties for the automatic concept extraction in the citation context, the system of automatic concept extraction is designed, which in limited cases, the system can automatically extract the concepts. The journal articles are used as the data source, and the result indicate that the system can fulfill the task of concept extraction in the citation context covering all the citing and cited documents

Keywords: Citation context; Concept extraction; Citation database; Citation index

目 录

图目录

表目录

引 言

题目背景及意义

在过去的一百年里，人类目睹了现代科技的高速发展和科技研究能力的飞速提升，而这二者的直接产物就是科技文献数量的急剧增长。据统计，PubMed 数据库生物医学领域科技文献的数量已经从 1986 年的七百余万篇增长到 2010 年的约两千余万篇，每年文献数量的复合增长率约为百分之四[1]，2005 年的文章发表数量为 666029 篇，平均每天文章发表数量超过 1800 篇[2]。科技文献不仅体现了科技工作者辛勤工作的成果，也构成了今后科技研究工作的基石。科研工作是在参考前人工作的基础上，对研究工作进行进一步展开拓展与创新。没有对前人工作的理解与借鉴，而贸然进行重复性的工作，将是对科研资源巨大的浪费。因此，对科技文献的合理利用，对于科研工作起到了至关重要的作用。然而，科技文献巨大的数量与急剧增长的速度无疑会给科技工作者查找相关信息带来的巨大的挑战。相较于以数据库形式进行存储的规则数据，科技文献的存在形式往往是非规则化的自然语言。为满足科技工作者的信息需求，信息科学工作者就必须提供更高技术含量、更加适应具体问题的信息抽取、信息检索和信息组织的工具与方法，从而应对自然语言所产生的复杂性。

科技文献是科技工作者工作当中最重要的信息载体之一。它区别于其它同样以自然语言形式存在的文档的重要特征在于科技文献包含参考文献。根据这一特点，加菲尔德建立了科技文献的引文索引系统。引文索引包括以顺序列表形式存在的参考文献以及这些参考文献的施引文献[3]。尽管引文索引所衍生出来的引文分析已经成为了文献计量学领域当中重要的组成部分，但是引文索引最初确实被它的发明者加菲尔德视为一种特殊形式的主题索引而应用于信息检索领域[4]。科技文献在对参考文献施加引用的同时，也会对参考文献当中的研究内容加以描述，如果可以接纳参考文献的主题与施引文献的主题具有相关性这一假设，那么引文链接关系的建立就表明了链接关系两端的参考文献和施引文献在主题上具有相关关系。到目前为止，包括汤姆森路透公司的 Web of Science、谷歌公司的谷歌学术搜索、微软公司的微软学术搜索在内的绝大多数科技文献检索系统都建立了引文索引，引文索引已经成为了科研工作者查找相关文献时必不可少的工具。

然而，引文索引相较于主题索引所具有的优点也成为了其进行信息检索时的障碍。引文索引避免了主题索引所存在的主题词不完备、不准确的缺点，

但却缺乏内容上的支持，使用者即使推定参考文献与施引文献之间具有相关关系，也无法通过引文索引来获知施引作者使用的到底是参考文献中的哪一个知识点，而这是缺乏内容支持的引文索引所无法解决的问题。能够克服这一问题的主要途径在于对引文上下文进行文本挖掘、信息抽取等相关工作。

在文献计量学领域，目前文献计量的最小单位往往是单篇的科技文献。通过对引文上下文当中的概念加以抽取，那么整个文献计量学就可以从以单篇科技文献为基本计量单位的阶段过渡到以概念作为基本计量单位的阶段。随着计算机技术的发展、科技文献可读文本化的实现以及科技文献开放获取运动的发展，对引文上下文的大规模计算机化处理已经成为了可能。因此，本研究所进行的引文上下文中的概念抽取工作就在这样的背景下展开。

本研究的意义在于通过对引文上下文当中的概念加以抽取，从而为科技文献信息检索当中通过引文索引进行检索的方式提供内容支持，同时也为文献计量学从以单篇文献为基本计量单位过渡到以概念作为基本计量单位提供帮助。

研究内容与方法

本研究着眼于引文上下文这一特殊的科技文献区域，使用自然语言处理方法自动地对名词性短语进行抽取以作为引文上下文当中概念的候选者，而后再根据名词性短语在施引文献当中的频率从而提取到相应的重要概念。

本研究使用的方法包括：

（1）文献调研法。通过大量阅读国内外文献，从而掌握目前相关研究的研究现状，搞清研究问题的来龙去脉，参考国内外学者的科研实践经历，为提出可行的研究方案打下坚实的基础。

（2）实验法。在提出研究方案的基础上，选取实际期刊文献作为实验数据，将设计完成的系统应用到实验数据上，并且最终得出实验结果，再对实验结果进行分析，从而最终得出结论。

本文的逻辑体系和结构安排

本文将首先描述关于本研究的相关工作，然后将描述引文上下文概念抽取的系统状况，而后，通过处理实际的科技文献得出实验结果并进行分析，最后进行讨论和总结。在第一部分相关研究工作的论述当中，将具体论述引文索引、引文上下文以及引文上下文概念抽取的相关研究。在第二部分系统设计与实现当中，将详细描述引文数据库的构建、引文句的提取、引文上下文中概念的抽取等过程。在第三部分，将描述实验数据和将实验数据应用在系统所得到的结果。最后，比较本文的研究与前人的研究，对研究当中的问

题进行讨论，并得出相关的结论。

1 相关研究工作

1.1 引文索引

几乎所有索引系统都拥有一致的基本目标，那就是为系统的使用者尽可能快速地提供他们所需要的信息。索引系统从传统的以人工标注方式为主的主题索引，发展到引文索引，再发展到目前商业搜索引擎复杂分布式自动化的索引系统，其背后的动机都在于协助使用者有效地在高速增长的海量信息中挑选出所需要的信息。因此，索引系统的设计者的主要目标就是对快速增长的海量信息加以合适的组织。

引文索引作为科技文献的特有产物，包括引文索引包括以顺序列表形式存在的参考文献以及这些参考文献的施引文献[3]。引文索引是加菲尔德受到夏帕达法律引文索引系统的启发，而为科技文献所构建的引文索引[5]。自 1873 年起，夏帕达公司就开始出版由法律专家提供的名为夏帕达引文索引的产品来为法律人士查找相关案例提供服务[6]。英国美国是以案例作为法律依据的国家，大量的法律文件都会指出某一判决是根据何种的案例而做出的。在这样的法律体系下，能否找到与当前官司相关的法律案例，对最终能否赢得官司起到了重要的作用，而夏帕达法律引文索引就成为协助法律人士查找相关案例的工具。

科技文献的引文索引与法律引文索引不同，是根据科技文献含有参考文献这一特点而发展起来的。从十九世纪开始，科学界逐渐形成了在文章中列出相关主题文章的科学规范。施加引用的作者会对所引用的参考文献加以描述，指出自己的文章受到了所引用的参考文献当中哪些知识点的启发。如果能够接受参考文献与施引文献之间在某一主题上具有关联这一假设，那么如果参考文献与施引文献之间存在链接关系，就能够揭示出参考文献与施引文献在内容上具有关联。由于一般而言，相较于主题标引员，施引文献的作者往往在参考文献和施引文献在主题的相似上有着更加清晰的判断，所以引文索引能够在协助使用者索取需求信息的过程中提供巨大的帮助。

在实际操作中，引文索引将参考文献与施引文献的引用关系记录下来，当使用者以一篇文献为入口进入引文索引系统之后，他就能够找到这篇文章全部的参考文献，以及在当前引文索引数据库当中全部的施引文献。与传统的以人工标注方式为主的主题索引不同，引文索引系统的设计者不必费尽周折地去选择适当的标引词来对文章加以标引，而只需要到文章底部参考文献列表的区域，将所有的参考文献加以记录即可。由于引文索引自身所具有的特点，它能够不依赖与术语学知识，同时也克服了信息索取者与主题标引者使用语言差异的困

难[4]。时至今日，包括汤姆森路透公司的 Web of Science、谷歌公司的谷歌学术搜索、微软公司的微软学术搜索等在内的绝大多数科技文献检索系统都在科技文献之间建立了引文索引，引文索引已经成为了科研工作者查找相关文献时必不可少的工具。

根据温斯多克的论断[7]，引文索引相比于传统的主题索引具有三个方面的优势。首先，引文索引克服了主题索引难以及时覆盖全部文献的缺点。其次，引文索引不存在主题索引缺乏连接跨领域、跨学科信息标注能力的问题。最后，引文索引避免了主题索引所遇到的语义学的困难。

对于温斯多克所提出的第一个优点，尽管今天相比于温斯多克的时代，索引系统已经发生了巨大的变化，计算机技术早已实现了文章全文的自动索引，主题索引已经呈现出新的形式。然而，新的索引系统仍然无法解决信息检索当中包括查询与文档不匹配等许多问题，这也就是 PubMed 数据库仍然使用医学主题词表对文章进行手工标记的原因。但是，每天发表的医学生物学文章多达近两千篇，可想而知，这对于标引人员是多么繁重的一项工作。相比较而言，引文索引的编制工作就轻松很多，并且也编制的过程也早已实现了计算机自动化，无论每天有多少新发表的科技文献，引文索引数据库都能够较为容易地实现更新。

针对第二个优点，引文索引可以更好地跨领域检索信息的原因还是在于施引文献的作者对于参考文献与施引文献的相关性上有着更加深刻的理解，而标引人员受限于标引数量以及自身的知识水平，往往难以标记出来跨领域的相关信息。引文索引天然地构成了两篇文章的关联关系，因此，也成为了跨领域、跨学科之间进行信息检索的重要工具。

就第三个优点而言，任何的标记语言都无法避免语义学的问题，能否通过所标记的词来完整的表达相对应的意思，实际上是很困难的一件事情。假如一个计算机程序仅仅从全文当中来对全部的内容进行索引，那么描述孟德尔定律的原始文章的索引结果就肯定不会出现孟德尔定律这样的词，因为，孟德尔本人并不会在文章中提出自己的研究成果称为孟德尔定律，但是，对这一研究成果加以使用的学者却都会称这一研究成果为孟德尔定律。这样，当使用者输入“孟德尔定律”这一检索词的时候，孟德尔定律的原始文章就不会出现在检索结果当中。相反，如果通过引文索引，在使用者找到含有“孟德尔定律”这个短语的文章后，他只需要找到这篇文章的参考文献，那么，孟德尔定律的原始文章就很可能出现参考文献的列表中。

如果主题索引是对文章内容的一种描述，那么引文索引就是文章外部特征的一种描述，尽管引文索引具有主题索引所不具有的优点，但是，不可否认的是，缺乏内容的支持成为了制约引文索引发展的重要因素。假设使用者希望通过引文索引来查找相关的信息，那么他就要首先通过关键词搜索的方法找到引文索引的起始文章。然而，在这篇起始文章的参考文献可能多达数十篇甚至上

百篇，而仅仅通过引文索引，使用者无法判断这数十篇甚至上百篇的参考文献和这篇起始文章有着怎样的关联，也不清楚这种关联是在何种程度上的，进一步说，使用者不清楚在这数十篇上百篇的文章当中到底哪一篇文章是与他的关键词相关，从而满足其信息需求的。因此，引文索引作为一种有效的信息检索手段，如果希望取得进一步的发展，更好地协助用户找到相关的信息，就必须有内容上的支持。

1.2 引文上下文

引文上下文是指科技文献当中，引文符号前后一个较小区域内的文本段。有学者将引文符号所在的句子称为引文句[8]，而沟通引文索引与内容相互联系的桥梁就在于对引文上下文进行深入的文本处理。怀特在综述当中指出[9]，信息科学领域针对引文上下文的研究包括对引文的性质进行分类、对引文上下文的内容进行分析和对施加引用人的动机分析。

引文性质的分类可以归结到一系列的分类框架当中，其中比较著名的是马尔维斯基和马热根森[10]分类体系，他们将引文分为四大类，而每一类又分为两类，这四大类分别是概念性引用和操作性引用、有组织的引用和敷衍的引用、扩展性引用与并列性引用、确认性引用与否认性引用。不同的作者对于引用的性质都有着不同的认识，因此，他们所构建的分类体系也有着较大的差异。引文上下文的内容分析是针对引文上下文的文本，手动或者自动地进行文本分析，从而应用于信息检索、文本摘要、文本理解等领域。引文动机的分析，可以理解为分析施引作者引用的原因。作者在引用某一篇参考文献的时候，往往有着不同的动机，这种动机可能是正面的，包括对研究的先驱者和相关的工作进行赞扬、标识出重要概念、提供背景资料等等[11]，动机也可能是较为负面的，例如提高恶意自引率等[12]。随着计算机技术的发展和学术论文可获得性的增加，引文上下文内容分析角度的研究正呈现着蓬勃发展的趋势

欧康纳[13, 14]较早提出利用计算机从引文上下文当中提取相关的检索主题词，来优化信息检索的结果。尽管欧康纳当时并没有机器可读的文档，但是他仍然提出了提取引文上下文并且挑选引文上下文当中检索词的实施步骤，并且将这些步骤实验于一组化学期刊文章[13]。除了挑选参考文献标记所在的句子，实验还根据一些手工操作的步骤将引文上下文进行扩展，从而尽可能的反映施引文献所对应的参考文献的特征。引文上下文当中的去除停用词的单词作为参考文献的特殊表示，被添加到原有参考文献的可检索信息（包括文献的人工检索词、题目和摘要中的词）当中。在文章里小规模实验的结果表明，文献检索的查全率从 50%提高到 70%。然而，在另一篇针对生物医药的文献中，文献检索检索的查全率仅仅提高了 4%[14]，欧康德将原因归结于生物医药领域平均每一篇参考文献的施引文献的数量较少。除此之外，斯莫尔[15]提出高被引文章可以被视为

概念象征，关于这方面的内容，会在接下来进行详细地阐述。

尽管早期科学家欧康纳、斯莫尔等人对引文上下文的内容分析进行了相关的研究，然而引文上下文的内容分析的研究并没有成为热点，在欧康纳、斯莫尔之后的十几年间一直处于较为停滞的状态。同时主要的工作都是基于人工对于内容的仔细阅读。主要原因大致有两个：首先，二十世纪九十年代以前机器可读科技文献还不普遍，大量的科技文献的主要存储格式是纸面形式，即便是以存储于计算机的文献，也无法实现大规模的文本化处理。二是互联网在当时还没有兴起，科技文献流通环节当中，文章的作者不能够像今天这样将文章上传到互联网上，出版商的产品也主要为纸面形式，研究者缺乏获取大规模全文文本资源的渠道。引文上下文与摘要不同，是根植在文章全文当中的。这样导致的结果就是科学家缺乏引文上下文研究的材料，同时的工作也不得不主要采用人工的方式。

这一状况随着互联网的兴起而逐渐改变。在互联网时代，科技工作者可以轻松地将自己的手稿上传到自己的个人网站或者类似于 arXiv 之类的个人文献自存档网站，同时，越来越多的出版商也逐渐将原有的单一化的纸面出版形式扩展到网络电子出版形式，而近些年逐渐流行起来的开放存取的出版形式，更使得只要是能够连入互联网的任何一个人都可以获取到科技文献原文，而不必付出任何费用。

网络科技文献数据库和搜索引擎 CiteSeer 在这样的背景下诞生了[16]。CiteSeer 通过网络爬虫搜集保存于互联网的科技文献，通过计算机程序对文献当中的内容进行处理，建立引文数据库。在文档的处理过程当中，CiteSeer 必须要完成准确定位参考文献列表，抽取参考文献信息，并且将参考文献与文章全文当中的参考文献标记相对应，从而找到参考文献对应的引文上下文。CiteSeer 采用启发式的方式来从参考文献列表的参考文献信息当中提取题目、作者、出版年、页数以及参考文献标记等信息。在实际操作过程当中，CiteSeer 需要通过使用正则表达式来应对参考文献标记因风格不同而呈现不同形式的问题。除了对于一篇文章内部的操作，CiteSeer 还要能够保证能够确定从不同文章当中提取的参考文献信息可以指向同一篇文章。由于参考文献信息的形式也是变化多样的，因此，Citeseer 采取的是参考文献信息统一化的方法，在这个方法当中要根据参考文献信息的长度进行排序，然后对参考文献信息子域当中的词和短语进行匹配。与以往文献数据库不同，CiteSeer 的一个重要特征就在于它提供了一篇文章的引文上下文，读者通过引文上下文就可以知道施引文献对于参考文献有着怎样的描述与评价。虽然目前 CiteSeer 已经被新版的 CiteSeer^{X} 所取代，但是，它仍然可以视为科技工作者开始通过计算机对引文上下文内容展开广泛研究的标志，在 CiteSeer 之后，针对引文上下文的研究蓬勃发展起来。

布拉德肖恩[17, 18]提出了称为参考文献直接索引的系统。该系统使用了从 CiteSeer 所提供的引文上下文当中提取索引词，再根据索引词和搜索词的相关性

和文献的被引次数对文章进行排名，这样，被引次数较多的文档就相对排名靠前。该系统取得了较好的效果，并且该系统揭示了引文上下文可以作为参考文献的一种特殊表达，将引文上下文当中的词语作为检索词添加到参考文献当中，可以在整体上提升信息检索的准确性。

里奇[19]同样从引文上下文当中提取检索词应用于信息检索，但是与布拉德肖恩不同的是，布拉德肖恩仅仅利用引文上下文当中的词语来作为参考文献的检索词，而里奇将引文上下文当中的词语与参考文献原文当中的词语组合在一起来进行信息检索工作。在研究中，里奇假定引文上下文含有对参考文献的有用的检索词。里奇的工作的主要贡献在于提出了用引文上下文当中的词语结合参考文献全文作为参考文献的表达形式，并且证明了将引文上下文当中的词语添加到参考文献的表达形式中时，确实有助于获得更好的检索效果。

阿加巴等人[20, 21]将引文上下文当中词语添加到参考文献的表达形式上，用于文档的分类和聚类。不仅如此，他们的工作还分析了这种混合文档表达形式的效果。在研究中，他们假定引文上下文可能会提供参考文献全文当中所没有的词以及参考文献中词语的近义词、上下位词以及其它相关词。在结果当中发现，引文上下文和参考文献全文的组合形式在聚类效果上超过了单纯的参考文献全文和基于链接的聚类效果，在分类上也提高了原有的分类效果。

难波等人[22-25]提出利用引文上下文的信息来完成一个特定领域的综述，系统的关键在于理清该领域文章之间的关系和从引文上下文当中抽取有用的信息。作者将引文分为三种类型，分别是引用他人理论或方法、与引用文章作对比或指出其缺点以及不属于前两种的其它类型。系统通过预先设置的提示词，来确定引文的类型。纳科夫等人[8]的工作集中于利用引文上下文的资源来管理愈来愈多的生命科学的文献。他们定义了一系列引文上下文的应用前景，包括作为未标记对比语料的数据源、同义词发现与排除歧义、参考文献的摘要、实体识别、关系抽取以及升级引文索引来提高文献检索准确率与回收率。在文章当中，纳科夫等人还将引文上下文视为对比语料来进行自动释义的工作。

图福等人[26, 27]在引文上下文当中引入标记的工作。他们根据句子的性质不同将篇章分析当中的论证区域的技术用来标记句子。工作的主要目的在于发现参考文章当中的创新点。实验结果表明，自动标记与人工标记一致性可以达到80%左右。艾肯斯等人[28]进行了数量化的分析，来揭示引入引文上下文对摘要以及信息检索的好处。特别地，他们分析了参考文献摘要和它的引文上下文之间的关系。他们的实验表明，引文上下文含有摘要当中所不具备的信息。

默罕默德等人[29]提出了利用引文上下文当中的词语产生多文档摘要的方法。相比于单纯利用参考文献原文的方法，引文上下文提供了额外并且在参考文献当中找不到的信息，这样就更加丰富了摘要当中的内容。

1.3 引文上下文的概念抽取

引文上下文中的概念的研究可以追溯到二十世纪七十年代斯莫尔[15]的工作。引文索引自诞生以来，就被加菲尔德视为一种特殊形式的主题索引[4]，某一个参考文献实际上是代表了其中的某一个主题或者概念。

然而，不同的人对一篇文章有不同的认识，从同一篇文章当中可能会获取不同的知识点。斯莫尔提出了高被引参考文献可以被视为概念象征的观点。他认为，在某一个参考文献的施引文献的不断积累过程当中，除了施引文献数量的增加以外，施引文献作者对于这一篇文章所引用的内容也不断趋向于一致，并且随着施引文献数量的不断增加，这种累积的效果将会越来越显著。因此，当一篇文章成为高被引的参考文献的时候，它就成为它所包含的被众多施引文献作者所引用的某一个主题或者概念的代表，或者说是概念的象征。根据斯莫尔的描述，引文上下文的概念可以指定义、方法、数据以及其它任何可以通过文本化描述的对象。

斯莫尔选取了 1972 年科学引文索引的 294 种化学期刊，并从当中提取出了被引次数最高的至少每篇 66 次被引的 52 篇文献。对于这 52 篇文献，每篇文献的施引文献当中随机挑选 12 篇，对于这 12 篇文献当中的引文上下文进行概括，计算相同的概括结果在 12 篇文献当中的分布。总体上， 52 篇参考文献的各自 12 篇施引文献有平均百分之八十七的施引文献描述的是同样一个概念。同时，在 52 篇参考文献当中，只有 10 篇参考文献的施引文献描述同一概念的比率低于百分之八十，并且这 10 篇文献当中有 7 篇文献的文献形式为书籍。书籍相比期刊文章，往往篇幅较大，也因此往往会覆盖较多的知识点。所以，相对而言，以书籍形式存在的参考文献的施引文献的作者引用的知识点也往往更加具有多样性。斯莫尔的结果表明了对于高被引参考文献，施引文献的作者在引用的内容上具有较高的趋同性。但是斯莫尔同时指出，他的实验结论是从被引次数极高的参考文献得来的，至于被引次数处于中等程度和较低被引的文献并不在他的文章的讨论范围里。这一重要结论构成了信息学领域共被引以及基于共被引的可视化发展的基石。尽管斯莫尔的工作具有极其重要的意义，但是，这一过程主要是通过手工进行完成的，工作需要研究者进行详细地阅读，另外也由于科技文献当时电子化的程度不普遍，在引文上下文当中的概念抽取工作趋于停滞状态。

施耐德[30]在斯莫尔工作的基础上，采取了半自动化的方式，工作的步骤主要分为三步。首先，通过文档共被引分析来建立概念集合。而后，通过软件工具将引文上下文当中的名词性短语抽取出来作为概念的候选者。最后，过滤出名词性短语作为概念。施耐德的工作主要是面向叙词的选取和使用名词性短语给文献聚类的结果添加标签，所以，他的研究当中的参考文献都是文献共被引

的文献，在所有参考文献当中这些文献具有最高数量的被引次数。然而，这种做法的缺点就在于忽视了全部不处于文献共被引当中、被引次数相对较低的文档。如果从施耐德挑选叙词和为文献聚类结果添加标签的角度而只挑选被引次数较高的参考文献作为研究对象是可行的。如果研究的目的是从引文上下文提取概念，那么只考虑在全部参考文献当中具有最高被引次数的一类参考文献就很难说是恰当的。另外，正如施耐德在文中所指出的，他的研究当中的概念抽取还处于半自动化的程度，这表示，参考文献和施引文献关系的确定、引文上下文的选取这些关键步骤都是手工进行。在将研究推广到大规模数量文献的要求下，就一定要保证引文上下文中概念抽取的每个步骤都能够实现自动化，尽量减少人工的影响，只有这样才能够真正实现大规模的概念抽取工作。

本文采用斯莫尔对于概念的观点，同时与施耐德的方法类似，将概念的范围缩小到以名词性短语形式存在的概念。从参考文献到引文上下文当中的概念的对应关系可以参见图 1.1。

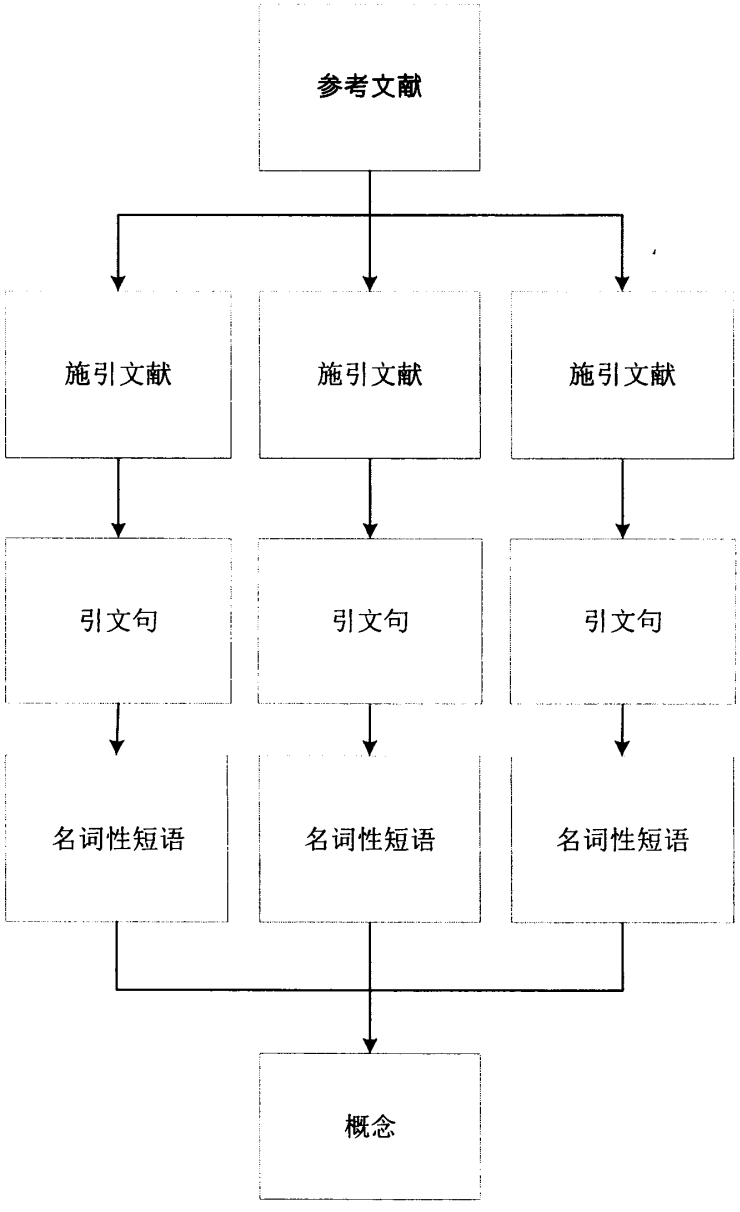

图1.1 参考文献到引文上下文当中概念的对应关系

2　引文数据库构建与参考文献匹配

为了实现从引文上下文当中提取概念的目标，系统要首先构建引文数据库以确定参考文献和施引文献之间的关系，而后将引文数据库当中的参考文献信息与期刊全文参考文献列表当中的参考文献信息进行匹配。在得到了正确的匹配结果之后，就能够确定参考文献符号在文章当中的位置，而这个位置也就是引文上下文的起点。

2.1 HistCite 软件与引文数据库构建

与 CiteSeer 所采取的方法不同，本研究中将采取通过 HistCite 软件来建立从科学引文索引数据库下载文献记录的局部数据库，再将局部数据库当中的记录和文章参考文献列表当中的记录加以匹配，从而建立与全文相联系的引文数据库。HistCite™ 是一套用来对科技文献直接引用关系进行分析和可视化的工具。程序的输入是汤姆森路透公司公司旗下 Web of Knowledge 产品格式的文献记录，输入要求这些文献记录在下载的时候要附带上每个文献的参考文献。如果这些输入的文献记录可以代表一个知识领域（例如一个特定领域在一个时间段的全部文章或者一个知名作者的全部作品），那么软件的使用者就能够对文献、期刊、作者、机构、关键的词汇加以排列。排列的原始依据有三个，分别为记录本身的个数，在所下载的文献内部所构成的局部引文数据库中的被引次数数据以及所下载的文献在 Web of Knowledge 下载时所选择数据库（科学引文索引、社会科学引文索引等等）中的被引次数数据。这里局部的含义是指该软件可以在所下载的带有参考文献的文献记录内部理清其中每一篇文献和其参考文献的引用关系，从而建立下载的文献记录内部的引文索引数据库。

从某种意义上讲，科学引文索引数据库也可以看成是一个局部的引文数据库，只不过这个数据库当中的记录个数极为庞大，覆盖了科学引文索引数据库所索引的期刊的全部文献记录，但是却不包含科学引文索引数据库之外的文献记录。HistCite 所构建的局部引文索引数据库相当于是在科学引文索引数据库中取出一部分的数据而构建的一个较小规模的数据库。其中每篇文献的局部被引次数是指在所下载的文献记录当中，这篇文章被引用的个数，可以得知局部被引次数一定是小于或者等于这篇文章在整个科学引文索引数据库当中的被引次数，因为只有包含了所有在科学引文索引数据库当中引用过这篇文章的文献记录的引文数据库才能够使得这篇文章的局部被引次数等于在科学引文索引数据库当中的全局被引次数，而如果有一部分没有被包含进去，那么局部被引次数就会小于全局被引次数。软件的输出是所研究知识领域下的不同形式的表格、

图形以及指标。在这个软件的帮助下，就能够对从科学引文索引数据库得到的记录进行可视化的分析，从而描绘出这一知识领域的发展脉络以及发展过程当中重要的文章[31]。

HistCite 除了作为对科技文献直接引用关系进行分析和可视化的工具，还有一个重要的功能就是该软件根据下载的带有参考文献的文献数据来构建局部的引文索引数据库[32]。构建的引文索引数据库是进行可视化分析的中间产物，它的操作步骤并没有明显的区别。在这个引文数据库当中，每一篇文献的参考文献都可以查找到，而每一篇参考文献的施引文献也可以查找到。对于不同的施引文献而拥有相同参考文献的情况，HistCite 会把这两篇文献都记录下来，最终在引文索引数据库里，不仅可以找到成对的参考文献与施引文献，也可以找到每个参考文献被多少篇施引文献所引用以及这些施引文献具体都是那些文献。在整个的处理过程中，对于数据的正确匹配显得格外重要。HistCite 这一软件产品属于汤姆森路透公司，它的开发者尤金加菲尔德，同时也是科学引文索引的发明者，因此，相比于其它的文献处理软件，对汤姆森路透公司科学引文索引等产品的数据格式，HistCite 具有较强的数据清洗和消除歧义的能力。科学引文索引的数据有很多的数据错误与不一致的情况，通过 HistCite 来协助建立引文索引数据库则能够在最大的程度上消除数据错误与不一致的情况，从而大大地减轻在建立引文索引数据库工作的负担。

2.2 参考文献匹配

在 CiteSeer 系统当中处理的对象直接就是文章当中的参考文献，所以工作的重点也就在于实现不同施引文献当中相同参考文献的匹配。由于在本研究中，采取了不同的方法，因此在通过 HistCite 以及科学引文索引下载的数据建立了引文索引数据库后，工作的任务就在于能够将引文索引数据库当中的参考文献的记录与 HTML 格式的期刊文章全文当中参考文献列表当中对应的参考文献记录进行匹配。

科学引文索引数据库当中，每一篇文献的参考文献的格式会根据该参考文献类型的不同而有所差异，但一般而言，对于每一篇参考文献，第一作者的姓氏以及名字首字母、文献出版年以及文献的来源这三个元素是必有的。文献的来源对于期刊文章就是期刊的名称，对于书籍而言就是书的名称，以此类推到会议论文等其他形式的文章。同时，对于期刊文章而言，参考文献的信息还会包括文章的卷数、页数以及 DOI（如果 DOI 存在的话）。在进行引文索引数据库当中的参考文献与期刊文章全文参考文献列表当中的参考文献的匹配之前，还应当从期刊文章全文当中，把每一篇参考文献的信息从 HTML 形式的源代码当中提取出来，这些信息包括每一篇参考文献的 DOI、参考文献的常规信息以及参考文献在 HTML 格式文章全文当中的编号。其中，由于有很多参考文献在

文章全文当中没有 DOI，所以程序只提取那些有 DOI 的参考文献的 DOI。参考文献的常规信息会根据期刊、出版社的不同而具有不同的风格，一般而言，都会包括全部作者的姓氏以及名字的首字母、文章的名称、文献来源的名称（对书籍而言，以书籍的名称来代替文章和文章来源的名称）、出版年等信息，对于期刊文章，往往还会包括期刊文章的卷数、期数以及页数。

参考文献在HTML格式文章全文当中的编号是本研究选择HTML格式文章的主要原因，通过这个编号，系统就可以直接地定位编号所对应的参考文献标记在文章主体部分的位置，从而确定引文上下文的起点位置。在 HTML 格式的文章当中，文章主体部分当中的每一个参考文献标记都含有一个超链接，读者可以点击这个超链接，再点击了之后，页面将会跳转到页面尾部参考文献列表的位置。其背后的原理是每一个超链接都有一个对应参考文献的唯一数字标识，换句话说，每一个数字标识都对应着参考文献列表当中的一个参考文献。因此，再进行完引文索引数据库和全文中的参考文献匹配后，系统就可以根据带有这个数字标识的 HTML 关键词准确定位参考文献标记在文章主体当中的位置。

提取的过程除了在HTML形式的源代码中通过查找每个信息的HTML关键词来定位该信息而后再进行抽取，还要做到将信息（主要是参考文献常规信息）当中的 HTML 标记去除、将参考文献常规信息中常见的 HTML 符号实体转换成对应的符号以及将以 ISO-8859-1 格式的字母转换成 ASCII 格式的字母以利于接下来的文本处理。尽管在本次研究所用到的科学计量学期刊的全部文章都是由英语来进行发表的，但是该期刊的 HTML 全文却是用 ISO-8859-1 来进行编码的，包括作者的姓名在内有大量西欧非英语字母的出现，如果不将这些字母加以转换的话，将会给接下来的文本处理带来诸多不便。因此，在本研究当中，将存在于 ISO-8859-1 字符集而不存在于 ASCII 字符集的字母加以转换。HTML 符号实体转换表可参考表 2.1。

表2.1 HTML 符号实体转换表

HTML 符号实体	实际符号
&#8211;	-
&#8212;	-
&#8216;	‘
&#8217;	‘
&#8220;	“
&#8221;	“
&amp;	&

由于引文索引数据库当中的施引文献的全文的文件名是按照“【卷号】.【起

始页】.html”的固定格式存储的，因此引文索引数据库可以较为容易地与含有文章全文的文件建立联系，接下来就是要将施引文献的参考文献与施引文献全文当中参考文献列表对应的参考文献相互匹配从而进一步找到参考文献唯一数字标识来定位引文上下文在全文当中的起点位置。

在理论上，DOI 是否相同可以作为两篇参考文献是否匹配的判断依据，在两篇参考文献都拥有 DOI 信息的情况下，如果 DOI 相同，则这两篇参考文献可以确定为是相同的参考文献；如果 DOI 不相同，则这两篇参考文献不是一篇文章。但是，在实践当中仍然会出现 DOI 错误的情况，那么，在这种情况下，DOI 就不能作为两篇参考文献是否匹配的依据了。如果在两篇参考文献当中有至少一篇没有 DOI，或者出现了 DOI 错误的情况，就必须使用其它信息来判断这两篇参考文献是否相同。

对于科学引文索引数据格式的文献记录，其中文献来源一项是以缩写形式呈现的，如果文献来源是期刊，那么记录当中就是期刊名称的缩写，如果文献来源是书籍，那么记录当中就是书籍名称的缩写，以此类推到会议论文等其它形式的文章。尽管，汤姆森路透 Web of Knowledge 的帮助文档中有期刊缩写与期刊全称的对应表格，但是，对于其中的书籍和会议，并没有这样的对应表。因此，对于本次研究而言，对于期刊论文只使用参考文献当中的第一作者姓氏及名字首字母、文献出版年、卷号以及页数，对于没有卷号及页数的文献（如书籍、会议论文），只使用第一作者姓氏及名字首字母、文献出版年。这样的判断方法实际上存在着错误的可能，当一篇施引文献拥有两篇第一作者相同、出版年相同的书籍参考文献，或者在更极端的情况下，当一篇施引文献拥有两篇第一作者相同、出版年相同、卷数和页数都相同的期刊文章参考文献的时候，判断都会出现问题。因此，当出现此种情况的时候，程序会做出提示，把判断的工作交由人工进行处理。

科学引文索引参考文献的目的在于让使用者能够通过参考文献当中的信息找到对应的文章，但是在利用这样的参考文献信息在进行匹配的时候，却因为文献来源是缩写形式，而带来巨大的困难。尽管汤姆森路透公司提供了期刊的对应表，但却没有给出每一个缩略单词和单词完整形式的对应表，同时，即使能够将每个缩写当中的词都找到对应的完整的英语单词，也无法完成匹配，因为类似于“of”一类的词也被省略掉了。在本研究当中的匹配方法并不完美，在一些情况下，还需要人工进行干预，但是，对于本研究当中的数据规模而言，考虑到完全进行所付出的代价，这种方法可以认为是适宜的。在实践的过程当中，并没有用到人工进行干预，可以推测，上文所提到的需要人工进行干预的情况，本身发生的概率也是较小的。

3　引文上下文的范围

3.1 英文文本的断句以及 Punkt 断句方法

在完成了引文索引数据库中的参考文献和全文参考文献列表当中参考文献的匹配之后，每一个参考文献就能够找到在全文当中的唯一数字标识，从而定位到该参考文献所对应的引文上下文的起始位置，接下来的主要工作就是找出引文上下文的范围。确定引文上下文的范围与本研究所做的在引文上下文中提取概念有着密不可分的联系。如果想要在引文上下文当中提取出所要的概念，就必须要确定引文上下文的范围，而引文上下文的范围这取决于所涉及到的重要概念在参考文献标记上下多大的一个范围内出现。可以看出这是一个两难的选择，过大的引文上下文范围很可能带来较多的噪音，从而抽取出一些完全不相关的概念，而过小的引文上下文范围有很可能无法抽取到所需要的概念，同时引文上下文的范围又在很大程度上取决于概念会在多大的范围内出现。

为了简化问题，在本次研究当中，只选取包含参考文献标记的句子进行研究，将引文上下文缩小为引文句。这样做的目的，就等于首先限定了一个变量，即引文上下文的范围，那么接下来就可以把工作的重点仅仅放在在包含参考文献标记的这个句子当中来选取名词性短语作为重要概念的候选者。事实上，英语文本的断句并不是一件容易的事情。表面上看，只要按照“.”、“?”、“!”三种符号就可以将这个句子断开，诸如，“Mr.”、“Washington D.C.”等词中也包含“.”，所以如果仅仅按照句子结束符号来进行断句的话，就会出现错误。这一问题在科技文献当中表现的尤其明显，例如表示页数的“p.”、“pp.”，表示举例的“e.g.”,表示“等等”的“etc.”等等。这些符号都对正确断句造成了极大的干扰。

在本次研究当中，采取 Punkt 断句方法，该方法是语言独立、无监督的断句方法。它是基于通过发现缩写的方式来去除这些缩写对断句所造成的影响。与以正确拼字作为线索的方法不同，该方法能够在只依赖三项标准的条件下高准确度地发现这些缩写。这三项针对缩写的标准是，缩写是由缩短的单词和结尾的句点所构成的紧密的语言搭配、缩写一般都很短、缩写有时候会在内部含有句点。该方法被应用于十一种不同的语言，并且还被应用到不同风格的语料上，取得了非常不错的效果。在实践中，该方法在处理科技文献当中的诸如“p.”、“pp.”、“etc.”、“e.g.”却效果不佳，因而，本研究在使用 Punkt 断句方法之前，就先把这些科技文档当中常常出现的缩写去除，然后再利用 Punkt 方法进行断句。

3.2 HTML 格式文本的断句处理

由于断句的方法必须要应用在包含有几个句子的一段话当中，因此，在获取了参考文献的唯一数字标识从而找到了引文上下文的起点后，还应当找到这个起点所在的段落。因为使用的是 HTML 源代码，所以应当以参考文献的唯一数字标识为起点，向上向下分别找到最近的<p>、 、<\tr>和<\p>、<\div>、<tr>标签来作为段落的起始位置和终止位置。同时，对于三个标签也要找到其中距离参考文献数字标识最近的标签。对于找不到上述标签的情况，就以当前文本的初始位置作为段落的初始位置、当前文本的终止位置作为段落的终止位置。值得一提的是<tr>、<\tr>标签，相比于<p>、 标签，它并不属于段落的分割符号，但是在网页中看，这个符号却可以被看做是段落之间的分割，所以要向前找到<\tr>的标签，向后找到<tr>从而在最小的程度上包含一个段落，从而有利于断句。应当注意到的是对于每一篇施引文献而言，它对同一篇参考文献可能在不同的地方进行多次地引用，因此就可能在一篇施引文献全文中出现多个引文上下文。

表3.1 引文句子示例

引文句子编号	引文句子
16.1	Due to its simplicity and meaningfulness, Hirsch's h index has created quite a stir in the scientific community.
16.2	The h index is defined as follows: A scientist has index h if h of his or her Np papers have at least h citations each and the other (Np h) papers have h citations each. It was calculated by sorting the hits according to times cited and counting manually or using the Citation report feature of the more recent version of the WoS.
23.1	The h-index has recently got attention and is assumed to be a robust measure for scientific performance and impact.
23.2	A scientist has index h if h of his or her NP papers have at least h citations each and the other (NP h) papers have h citations each. .
27.1	Some of the most commonly used indicators to measure the scientific output of researchers that we can find in the literature are :
27.2	In the last few years, the scientific community has paid a lot of attention to a new index, introduced by Hirsch 2005 and called the h-index.
27.3	The h-index was originally presented by Hirsch.
29.1	This paper offers three axiomatic characterizations of the Hirsch index; see Wikipedia for a discussion of advantages and criticisms of the Hirsch index.

不仅如此，在选取的最小段落当中甚至是一句话当中，也可能出现多个参考文献标记，因此，在实际的处理过程当中一定要注意这些问题。在断句的过程当中，首先根据起始点和终结点将所提取的段落文字进行断句，在每一句话当中，如果发现所需要的参考文献的数字标识，就把这句话提取出来，同时把新的段落起点设置为该句的句子尾部，从而保证同一句当中的参考文献数字标识不会干扰到接下来的断句，同时也保证了同一个段落里的其它的引文句的抽取不会出现问题。在抽取了含有数字标识的引文句子后，要对句子进行清理，把句子当中 HTML 标签去除，同时把句子当中的参考文献标记全部去除。

在本研究所选取的实验数据当中，参考文献标记是较为常见的作者加上出版年格式，要注意到去除参考文献标记不仅会去除掉所需要的参考文献的参考文献标记，也会把句子当中其它的参考文献的参考文献标记同时去除。去除参考文献标记带来的好处就是，参考文献标记中的作者名字和出版年都会在同一个参考文献的不同施引文献的引文上下文当中反复出现的，去除参考文献标记以后，程序就不会提取参考文献标记当中的作者名字和出版年作为重要概念了。提取出来的引文句子示例可以参见表 3.1，表格当中的八句话都是针对同一篇参考文献的，16、 23、27、29 这些数字表示这些是不同的施引文献，可以看到，对于编号为 16 的这篇施引文献当中对于参考文献拥有两个不同的引文句，而编号为 27 的这篇施引文献当中对于参考文献则拥有三个不同的引文句，编号为 29 的这篇施引文献当中对于参考文献则只拥有一个引文句。

4 概念的抽取

4.1 英文文本的分词

在完成了提取参考文献唯一数字标识（或者说参考文献的引用标记）所在的句子来作为引文上下文之后，接下来的处理过程就进入到了词的层面上。首先需要完成的任务就是句子当中词与词之间的分离。汉语当中的词与英语中的词有很大的区别，汉语的词和词之间并没有空格，所以汉语的分词是在中文文本处理过程当中很重要的一个步骤。然而，即使英语当中的词之间有空格的情况下，也不能够仅仅依赖空格来对句子当中的英语单词进行分离。举例而言，如果仅仅使用空格来对句子进行分割的话，那么，分割出来的结果就会出现“(This”、“fine.”一样的分割结果，很显然这样的分割结果是错误的，正确的分割结果应当是“This”和“fine”，所以，在对英语单词进行分割的过程当中，也需要慎重地对待，词的正确分割对接下来的处理有着非常重要的意义，因为，接下来的步骤都会用到分割好的词，如果词的分割不合理，那么就会造成余下文本操作结果的错误。一些早些的文本信息处理系统将词定义为超过三个字母的连续字母数字文本序列，并且所有的大写形式都会转换成小写形式，这些文本序列之间以空格隔开[33]。尽管这一种较为简单的分割方法可以适应小规模的语料，但是，对于信息抽取任务，就尽量保留词的原貌，从而尽可能多地保留更多的信息。在实践的操作过程当中，我们采用宾夕法尼亚大学自然语言处理 Treebank 项目所提供的分词规范对引文句当中的单词进行分割[34]。这些规范包括：大部分的标点都从和它们相连的单词上分割出来，双引号将被两个单引号，动词缩略形式与盎格鲁——撒克逊式名词所属格将被分解成它们的组成词素，并且每一个词素在接下来的处理过程中将会被分别进行标注。其中动词缩略形式与盎格鲁——撒克逊式名词所属格将被分解成它们的组成词素的例子是 children's 转化为 children 's、parents' 转化为 parents '、can't 转化为 ca n't、gonna 转化为 gon na。这个单词分解器假设文本已经被分解为句子，因此除了句子当中的最后一个单词和它的句点会分离，句子当中的句点都跟随句点前面的单词，比如 e.g. 就被视为独立的一个单词，从这里我们就可以看到前面的断句对于后来的工作也起到了极其重要的作用，如果断句错误，造成了一句话中实际上包含了两个句子，那么，第一个句子的最后一个单词就会被错误地分割出来，这个单词的末尾就会多出一个多余的句点。在具体的程序操作当中，程序首先进行动词缩略形式与盎格鲁——撒克逊式名词所属格词素的分解，然后再将除了“.'-/,&_”之外的标点符号作为分割单词的分割符号，接下来再把后面不接空格的逗号和

单引号分割出来（2,500 这个逗号不会分割，因为逗号后面跟随的不是空格），然后把新句子之前或者字符串末尾的句点分割出来，最后再将整个字符串通过空格来进行分割，这样在引文句当中分割单词的任务就完成了。在最后的分割结果当中，长度较短的单词被保留下来，连字符号得以保留同时由连字符连接起来的多个单词视为一个单词，单词当中的大写字母不进行小写化处理。

4.2 词性标注

每一个所提取出来的单词会被赋以相应的词性。词性标注是给语料库当中的每一个单词赋以词性分类的过程。由于词性标注的过程要首先去除标点符号，因此，上一步所进行的词与词之间的分割就显得极为必要。对英语而言，词的种类可以首先分为封闭种类的词和开放种类的词，而开放种类的词又主要可以分为四类，名词、动词、形容词和副词[35]。在这里，封闭种类的词当中的元素相对固定，而开放种类的词可以动态地进行扩展。名词主要是用来表示概念，在这里概念的含义较广，可以指人、地点、事件、事物等等，例如“中国”、“小明”、“九一八事变”等等。在英语当中，名词之前往往有限定词。动词主要是用来描述动作和过程的，例如“打”、“骂”、“哭”、“笑”。形容词主要是用来描述名词，例如“美”、“丑”、“善”、“恶”都是用来描述人的。副词可以用来描述动词，也可以用来描述形容词。封闭种类的词，包括介词、人称指代词、冠词等等。在词性标注的算法当中，输入是一串的词语和特定的词性集合，输出是每一个词都附带了最合适的词性。大多数的词性标注方法都是基于规则和统计的。在本次研究当中，采取的是统计词性标注方法当中的最大熵词性标注方法，该词性标注方法是通过宾夕法尼亚大学 Treebank 词性标注语料训练而成的[36]。

4.3 名词性短语与概念抽取

根据每个单词在文本当中所显示的词性，就可以进一步将文本当中的名词性短语提取出来。名词性短语的提取过程就是在文本当中提取与名词性短语相对应的文本片段。在名词性短语的提取过程当中，最重要的资源之一就是前面步骤所完成的单词的词性，通过对单词词性的处理，将大大有利于名词性短语的提取。在语料当中，名词性短语最常见的标注方法就是 IOB 标记，其中 I 表示单词在名词性短语之中，O 表示单词在名词性短语之外，B 表示单词是一个名词性短语的开始。应当注意到，只通过 B-NP 和 I-NP 两种标记也可以表示名词性短语，对于那些不在名词性短语当中的单词，可以通过 O 来进行表示，不进行标注也不会产生歧义。

与词性标注的相似，名词性短语的提取通常也可以大致分为基于规则和基

于统计的方法。在本次研究当中，采取的方法是基于统计方法当中的朴素贝叶斯方法来提取文本当中的名词性短语。语料是基于拥有二十七万单词规模的“华尔街日报”标记文本的 CoNLL-2000 语料。在基于预料当中前百分之九十作为训练语料，后百分之十作为测试语料的评估中，该方法的 F 值可以达到百分之九十左右。要注意到的是每篇施引文献针对同一篇参考文献的引文上下文可能有多个（在本研究中为引文句）。因此，在每一个引文句进行名词性短语的提取结束之后，需要把施引文献针对一篇参考文献的全部名词性短语整合到一起。表 4.1 展示的是一篇参考文献的三篇施引文献的所有针对这篇参考文献的引文句当中全部名词性短语。

表4.1 三篇施引文献的全部名词性短语

施引文献 A	施引文献 B	施引文献 C
h index	Some	paper
attention	most	three axiomatic characterization
robust measure	used indicator	Hirsch index
scientific performance and impact	scientific output	Wikipedia
scientist	researcher	discussion
index h	that	advantage
h	we	criticism
his	literature	
NP paper	last few year	
least h citation	scientific community	
each	lot	
other NP h	attention	
paper	new index	
h citation each	Hirsch 2005	
	h index	
	Hirsch	

一篇参考文献所对应的引文上下文当中的所有名词性短语出现在这篇参考文献的施引文献的个数是有差异，而那些出现在较多施引文献的名词性短语无疑要比那些出现在较少施引文献的名词性短语有着更高的重要性。不同施引文献的作者在对参考文献的描述当中，使用了相同的名词性短语，这就意味着该名词性短语的固定形式已经受到了不同施引文献的作者的认同，那么，该名词性短语就已经从字符串上升到了概念的层次上。

系统将把出现在不同施引文献个数较多的名词性短语提取出来作为概念。在

将高频次名词性短语作为概念提取之后，系统还会考虑包含每篇参考文献对应的拥有最高频次的名词性短语的名词性短语，以此作为概念文本表达形式的扩展形式进行进一步的分析。

5 系统实验

5.1 实验数据

实验数据是科学计量学（Scientometrics）期刊 2010 年到 2011 年全部 HTML 格式的全文。每个文件以“【卷号】.【起始页】.html”的格式进行存储。数据是在 2011 年 12 月当该期刊实施短暂开放存取的时候下载的。所有的这些文章是以 HTML 形式来进行存储的，相较于以 PDF 形式存储的文件，在文本的处理上，会带来非常大的便利。同时，由于数据是以 HTML 形式来进行存储的，HTML 标签可以用来进行对参考文献标记的准确定位，这是 PDF 和 TXT 形式的纯文本文件所不具备的特征。由于该期刊在 2010 年和 2011 年都是被汤姆森路透公司的科学引文索引检索系统收录的，因此，所有的期刊文章都能够在科学引文索引数据库中找到对应的记录。在下载全部的全文数据的同时，也将科学计量学期刊在科学引文索引数据库当中 2010 年到 2011 年的记录全部下载下来，以 TXT 文件的形式进行保存。在去除了两篇没有参考文献的编辑社论后，又发现在期刊全文当中实际上有一篇文章实际上并没有以 HTML 格式存储，这篇文章只有 PDF 格式的文件。在整理完毕之后，得到了 455 篇以 HTML 格式存储的期刊文章。

科学计量学期刊强调以定量的方法对科学发展和科研机制进行探索，主要的研究范围就如期刊的名称一样，是科学计量学。在实践当中，科学计量学通常以文献计量学作为其重要工具来衡量科技文献的影响力。研究的方法包括定性和定量，但是强调以统计学的定量方法来对科学进行研究。该期刊发表科学计量学领域的原创研究成果、短篇交流、基础报告、综述文章以及书籍评论等内容。由于该期刊广泛的覆盖面，该期刊成为了科研工作者和科研管理人员必不可少的信息来源。该期刊为在科研机构工作的图书馆员和信息工作者提供了宝贵的信息支持。该期刊自 1978 年创刊以来，就成为了科学计量学领域最为权威的期刊之一。选取该期刊的主要目的也在于本研究团队有科学计量学领域的专家，他们对该领域近些年的发展状况，特别是经典的和新出现的领域知识有着深刻的理解，可以对实验的结果进行人工上的检验。

在数据的获取上，本研究采取了使用计算机程序进行自动获取的方式。对于每一刊的期刊文章都有一个对应的网页，在网页中会列出全部的出版在这一刊的期刊文章。对应的网页的链接地址中含有“【期刊纸版 ISSN】/【卷号】/【刊号】”这样的固定格式。因此，通过遍历对应的卷号和刊号，就能够找到所需要出版年份的所有期刊文章。在每一刊对应的网页中，可以通过固定的

HTML 关键词来定位查找到每一篇文章的链接，这样就能够把该链接所指向的 HTML 文件下载下来了。在本研究所涉及到的 2010 年到 2011 年的期刊全文，对应着该期刊第八十二期第一刊至第八十九期第三刊，其中每一期都有三个刊的文章。在这种通过 HTML 关键词来定位信息的程序中，要保证的一点就是 HTML 关键词的数量一定要和所需要信息的数量保持一致，在本研究当中，所需要的信息指的就是每一篇期刊文章全文所在的网页地址。

在通过汤姆森路透公司所下载的引文数据当中，尽管 HistCite 软件本身就拥有很好的数据清理能力，然而，由于数据本身的错误和不一致现象，所以为了实现该数据的最终可用性，本次研究还是在数据清理上费了一番周折。应当指出的是，HistCite 软件并不能够处理在第一作者、出版年、出版来源、期号相同，而 DOI 不相同的情况。表 5.1 显示的是 Newman 在 2001 年发表于物理评论 E 上的三篇文章，与一般的期刊文章不同，这三篇文章并没有标明它们对应的在期刊当中的页数而是通过 DOI 来进行区分，但遗憾的是，HistCite 显然并不是通过 DOI 的不同对不同的文章进行区分的，在这两篇文章的第一作者、出版年、出版来源、期号相同的情况下，HistCite 软件错误地将这两篇文章视为同一篇文章，可以看出 HistCite 在判断两篇参考文献是否为同一篇文章的时候根据的是这两篇参考文献是否拥有同样的第一作者、出版年、出版来源、期号和页数。与此同时，由于引文数据当中自身的错误，也往往会导致 HistCite 所建立的引文数据库当中出现错误，例如，表 5.2 所显示的两篇参考文献，它们在同一篇施引文献的参考文献列表当中同时列出。

表5.1 Newman 的三篇文章

Newman MEJ, 2001, PHYS REV E, V64, DOI 10.1103/PhysRevE.64.016131
Newman MEJ, 2001, PHYS REV E, V64, DOI 10.1103/PhysRevE.64.016132
Newman MEJ, 2001, PHYS REV E, V64, DOI 10.1103/PhysRevE.64.026118

表5.2 同一篇参考文献不同信息

Moed HF, 2008, SCIENTOMETRICS, V74, P153, DOI 10.1007/s11192-008-0108-1
Moed HF, 2008, SCIENTOMETRICS, V74, P153, DOI 10.1007/s11192-008-0108-1, 2008, RES ASSESSMENT EXERC

然而，可以看到，根据第一作者、出版年、出版来源、期号、页数、DOI 等信息判断，这两篇参考文献是同一篇文章，只不过在第二篇文章末尾又多出了一些信息。由于一篇施引文献的同一篇参考文献不能够列出两次，很显然这是科学引文索引数据库在数据录入时所出现的错误，这一错误带来的结果就是在通过 HistCite 软件所构建的引文数据库当中，对应那篇参考文献的施引文献

会多出一篇来。也就是说，HistCite 还是通过第一作者、出版年、出版来源、期号、页数来区分两篇文章，尽管在表 3.2 当中，第二篇参考文献比第一篇参考文献多出了一些信息，但是，HistCite 正确地把它们视为同一篇参考文献，所带来的问题就是这两篇参考文献所对应的施引文献被二次记数，对应与那篇参考文献，在总的施引文献数量上会多出一篇。但是，幸运的是，在详细查看施引文献的信息的时候，实际的施引文献数量会和引文数据库当中的数量出现不一致的情况，所以，同一篇参考文献多次出现在一篇施引文献参考文献列表当中所造成的最终施引文献记数错误的问题是可以通过程序进行发现的。

在处理完以上两个引文数据库当中的数据问题之后，最终根据在科学引文索引上所下载 455 条对应 HTML 格式期刊文章全文的引文数据，而构建了这 455 篇文章所构建的引文数据库。在这个引文数据库当中，根据 HistCite 的统计结果，总共有 842 位作者，9117 篇参考文献，平均每一篇施引文献对应大约 20 篇的参考文献，题目当中去除停用词的单个词共有 1355 个。在 9117 篇参考文献当中，本次研究选取被引次数在五次及五次以上的参考文献作为研究对象，这样的参考文献总共有 263 篇，这样就保证了对于这 263 篇参考文献当中的每一篇文献都对应着至少五篇施引文献，同时也对应着至少五个引文上下文。应当指出的是，在这 263 篇参考文献当中，没有一篇参考文献同时也是 455 篇施引文献当中的文献。在 455 篇施引文献当中，有六篇文章被其它的施引文献引用达到五次及五次以上，而在本次研究当中为了简化问题，没有将这六篇文章作为研究对象，这样就保证了 263 篇至少五次被引次数的参考文献和 455 篇施引文献的集合并不存在交集。在 263 篇参考文献当中，最高文献的被引次数达到了 86 次，是 Hirsch 关于 h 指数的开山之作，近些年来 h 指数一直是科学计量学领域的研究热点，在全部 455 篇施引文献当中有 18.9%的文章引用了这篇文章。与此相比，第二高被引的引用次数则急剧下降到 28 次，可见 Hirsch 关于 h 指数的这篇文章在科学计量学领域的影响力之大。在 263 篇参考文献里，平均每篇文章的被引次数为 7.97 次，被引次数的中位数是 6，标准方差是 6.073，10 次以上被引次数的参考文献有 46 篇，占全部 263 篇参考文献的 17.5%，10 次及 10 次以下的参考文献有 217 篇，占全部 263 篇参考文献的 82.5%。263 篇参考文献当中，不同被引次数的参考文献对应的频数与百分率的详细信息可以参见表 5.3，参考文献被引次数的统计信息可以参见表 5.4。

表5.3 不同被引次数的参考文献对应的频数与百分率

被引次数	频数	百分率	有效百分率	累积百分率
5	91	34.60	34.60	34.60
6	56	21.29	21.29	55.89
7	25	9.51	9.51	65.40

表 5.3 不同被引次数的参考文献对应的频数与百分率（续）

被引次数	频数	百分率	有效百分率	累积百分率
8	22	8.37	8.37	73.76
9	10	3.80	3.80	77.57
10	13	4.94	4.94	82.51
11	9	3.42	3.42	85.93
12	9	3.42	3.42	89.35
13	5	1.90	1.90	91.25
14	6	2.28	2.28	93.54
15	4	1.52	1.52	95.06
16	3	1.14	1.14	96.20
17	2	0.76	0.76	96.96
18	1	0.38	0.38	97.34
19	2	0.76	0.76	98.10
21	1	0.38	0.38	98.48
22	1	0.38	0.38	98.86
23	1	0.38	0.38	99.24
28	1	0.38	0.38	99.62
86	1	0.38	0.38	100.00
总共	263	100	100	

表5.4 参考文献被引次数的统计信息

项目	数值
平均值	7.97
中位数	6.00
众数	5.00
标准差	6.07
方差	36.88
范围	81.00
最小值	5.00
最大值	86.00

在引文数据库和文章全文参考文献的匹配过程当中发现，凡是以“10.1556/SCIENT”开头的 DOI 都会出现 DOI 找不到匹配项的情况，通过 DOI 服务也找不到这些以“10.1556/SCIENT”开头的 DOI，可以判断可能是这些文

章的 DOI 已经和数据录入科学引文索引时候相比发生了改变，因此以“10.1556/SCIENT”开头的 DOI 不作为参考文献是否相同的判别依据。在匹配的过程当中，也会出现很多数据错误或者不一致的现象。有的施引文献在参考文献当中列出作者的姓氏和教名中间名的首字母，而有的施引文献则在参考文献当中只列出作者的姓氏和教名的首字母而不列出中间名的首字母，如“Hirsch, J. E.”和“Hirsch, J.”、“King, D. A.”和“King, D.”。有的参考文献当中出版年份出现错误，如 2007 年写成 2003 年。有的参考文献当中作者名字写错，如“Bassecoulard, E.”写成“Basselcoulard, E.”。有的期刊两年只出了一期的文章，因此这篇参考文献实际上拥有两个出版年。还有一种是作者名字的不一致，例如“Price, D.J.D.”被写成“de Solla Price, D.J.”，“De Nooy, W.”被写成“Denooy W.”，“Van Raan, A.F.J.”被写成“VanRaan, A.F.J.”，实际上，这几种名字的写法都没有错误，只不过是在形式上不同而已。在本次研究所进行的参考文献匹配过程当中，是将科学引文索引当中的参考文献与实际期刊文章全文参考文献列表当中的参考文献进行匹配，在科学引文索引的数据录入的过程当中，数据本身就进行了相应的清洗与排除歧义，而如果采取于类似 CiteSeer 的方法，就必须做到通过程序对数据进行自动的匹配，例如，“Price, D.J.D.”要能够被认定为与“de Solla Price, D.J.”是一致的，可以想象这一过程绝不是简单直接的。

5.2 实验结果

在最终获得名词性短语个数的统计结果中（参见表 5.5），263 篇参考文献平均每篇参考文献的所有引文上下文（引文句）当中有 83.92 个不重复的名词性短语，名词性短语最多的参考文献包括有 751 个名词性短语，而最少的参考文献只有 20 个名词性短语，而这两篇参考文献则分别对应着被引次数数量最多、达到 86 次的 Hirsch 的 h 指数文章和被引次数最少、为 5 次的一篇文章。被引次数越多就意味着施引文献越多，而施引文献的次数则大致上决定了引文上下文的数量，更多的被引次数因而往往会带来更多的引文上下文和更多的名词性短语。被引次数较多的参考文献因而在提取的名词性短语的数量上往往超过被引次数较少的参考文献。根据这种情况，将每一篇参考文献的名词性短语的数量除以这篇参考文献的施引文献的数量，从而得到每一篇参考文献的平均每一篇施引文献的名词性短语的数量。可以看出，每一篇参考文献的名词性短语的数量除以该参考文献施的引文献数量的平均值是 10.82，也就是说平均每篇施引文献的所有引文句（一篇施引文献可能有多个引文句）当中会出现 10.82 个名词性短语。同时可以看到，统计结果当中的方差和标准差都有了大幅度的下降，表示平均每篇施引文献的所有引文句当中名词性短语的数量更加趋向于平均值。在全部 263 篇参考文献当中，参考文献的名词性短语的数量除以该参考文献的施引文献数量的最大值是 30.6，而最小值是 3.13，而拥有最大值的参考文献的

施引文献的数量只有5，但是，它的每一篇施引文献却拥有高达4.6句的引文句，可见，影响参考文献的名词性短语的数量除以该参考文献施引文献数量的值的大小的因素不再是施引文献的数量，而是平均每篇施引文献当中的施引句的数量，施引句子越多，所包含的名词性短语越多的可能性就越大。在进行了全部参考文献的名词性短语的统计之后，接下来关注的是那些被引次数大于10次的46篇参考文献。由于被引次数较多的参考文献往往在提取的名词性短语的数量上超过被引次数较少的参考文献，所以可以解释在平均值的数量上，被引次数大于10次的参考文献的数量大于全部参考文献数量的现象。然而，在参考文献总的名词性短语的数量除以该参考文献施引文献数量的值的统计结果中，平均值都小于以全部参考文献作为研究对象的统计结果当中的平均值，其中平均值为9.99小于以全部参考文献作为研究对象时候的10.82，中位数是9.78小于以全部参考文献作为研究对象时候的10.2，同时还可以发现，被引次数超过10次的参考文献的统计结果的标准差和方差也小于以全部参考文献作为研究对象时统计结果的标注差和方差，其中标准差为2.62小于以全部参考文献作为研究对象时的3.89，方差为6.86小于以全部参考文献作为研究对象时的15.14。这种情况表明，对于被引次数超过10次的参考文献，尽管在总的名词性短语的数量上由于被引次数较高而远远大于被引次数较低的参考文献，但是，在参考文献总的名词性短语的数量除以该参考文献施引文献数量的值上，总体上却小于被引次数较低的参考文献，同时，被引次数超过10次的参考文献在参考文献总的名词性短语的数量除以该参考文献施引文献数量的值上也更加趋向于平均值。

表5.5 参考文献对应名词性短语个数的统计结果

项目	A	B	C	D
平均值	83.92	10.82	159.93	9.99
中位数	73.00	10.20	129.00	9.78
众数	54.00	10.80	95.00	5.69
标准差	61.42	3.89	108.70	2.62
方差	3772.54	15.14	11816.51	6.86
范围	731.00	27.48	678.00	10.15
最小值	20.00	3.13	73.00	5.69
最大值	751.00	30.60	751.00	15.85

A: 名词性短语数量（全部参考文献） B: 参考文献总的名词性短语的数量除以该参考文献施引文献数量的值（全部参考文献） C: 名词性短语数量（被引次数大于10的参考文献） D: 参考文献总的名词性短语的数量除以该参考文献施引文献数量的值（被引次数大于10的参考文献）

接下来的统计结果针对名词性短语至少出现在两篇不同的施引文献的参考文献的个数。由于在本研究当中是把引文上下文中提取的名词性短语作为概念表达词的候选者，而且通过判断这个候选者是否在多个文献当中共同出现，以及共现的频数频率作为依据来从候选的名词性短语当中挑选出表示概念的词，所以，已经挑选出来的名词性短语，能否被“选”为代表概念的词就要首先满足在同一篇参考文献的至少两篇不同的施引文献的引文上下文当中出现的条件。在确定名词性短语的状况之前，首先要确定的是对应名词性短语的参考文献的统计状况，表 5.6 显示的是在满足参考文献对应的引文上下文中的同一个名词性短语至少在两篇不同的施引文献引文上下文出现的前提条件下，同时存在对应名词性短语出现在不同百分率及以上的施引文献的参考文献的数量。

表5.6 不同百分数条件下存在对应名词性短语的参考文献的数量

百分数	参考文献的数量（全部）	参考文献的数量（被引次数>10）
0	248	44
10	248	44
20	247	43
30	217	27
40	171	21
50	77	9
60	63	6
70	19	1
80	10	0
90	3	0
100	2	0

比如，在表 5.6 中百分之五十及以上的参考文献的数量为 77，这表示的是这77篇参考文献当中的每一篇参考文献所对应的引文上下文中存在名词性短语出现在这篇参考文献所对应的百分之五十及以上的施引文献当中。如果一篇参考文献满足百分之五十的条件，同时这篇参考文献有 5 篇施引文献的话，那么，对应的参考文献的名词性短语就要出现在参考文献所对应的至少三篇施引文献当中（3>5/2=2.5>2）。图 3.1 对应表 5.6，纵坐标代表的是参考文献的数量，横坐标代表的是施引文献的百分数。从表 5.6 中可以看到，在百分数为 0 的时候，也就是只要满足参考文献对应的引文上下文中名词性短语至少出现在两篇不同施引文献引文上下文的条件，而不受到要超出对应施引文献百分率限制的时候，在 263 篇参考文献当中，总共有 248 篇参考文献满足该条件，而其它的 15 篇参考文献对应的引文上下文中没有名词性短语出现在参考文献的两个及两个以上

的施引文献当中。从图 5.1 中可以看出，参考文献数量下降速度最快的是百分之四十到百分之五十这个区间，在百分之五十以上的区间，能够满足参考文献所对应的引文上下文中的名词性短语出现在这篇参考文献所对应的百分之六十及以上的施引文的参考文献的数量有 63 篇，能够满足参考文献所对应的引文上下文中的名词性短语出现在这篇参考文献所对应的百分之八十及以上的施引文的参考文献的数量有 10 篇，而能够满足参考文献所对应的引文上下文中的名词性短语出现在这篇参考文献所对应的百分之九十及以上的施引文的参考文献的数量有 3 篇，最后还是有两篇参考文献中对应的名词性短语对出现在这两篇参考文献所对应的全部的施引文献的引文上下文当中。这两篇参考文献的施引文献的数量都只有五个，是所挑选的参考文献施引文献数量的下限。

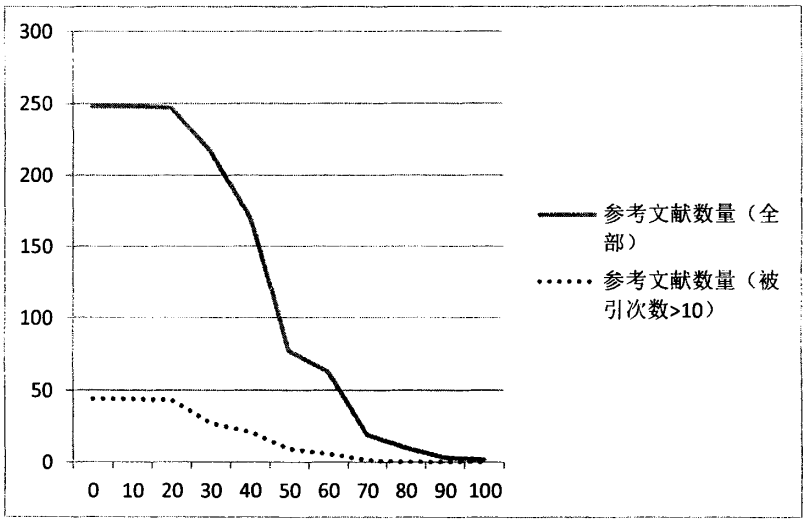

图5.1 不同百分数条件下存在对应名词性短语的参考文献的数量

在不同的百分数之下全部参考文献的每篇参考文献对应的引文上下文当中的名词性短语个数的统计结果可以参见表 5.7 和表 5.8。以全部参考文献作为研究对象的时候，在初始阶段百分数为 0，即只要求参考文献对应的引文上下文当中的名词性短语满足出现在参考文献的两篇及以上的施引文献的引文上下文里，可以看到，平均每篇参考文献里，满足条件的名词性短语有 6.92 个，正如前面所提到的，在这种情况下，全部的 263 篇参考文献当中，有 15 篇参考文献没有满足条件的名词性短语（存在名词性短语出现在两篇及两篇以上的施引文献当中），而其它的 248 篇参考文献有满足条件的名词性短语，由于那些没有满足条件的参考文献对应的名词性短语的数量都是 0，所以会从整体上降低名

词性短语数量的平均值。在百分数达到四十的时候，即参考文献的名词性短语要出现在该参考文献对应的百分之四十及以上的施引文献当中的情况时，统计结果当中的众数为 0，此时的平均值为 1.76，中位数为 1。在百分数增加到为五十的情况下，平均值就已经降低为 0.4，而中位数和众数则降低为 0。在百分之五十以上的情况中，平均值逐步降低，最终达到在百分之百时候的 0.01，而此时，满足条件的参考文献有两个。

表5.7 参考文献对应的名词性短语个数（0%-40%）

	0	10	20	30	40
平均值	6.92	6.27	4.59	2.93	1.76
中位数	5.00	5.00	4.00	2.00	1.00
众数	2.00	2.00	2.00	1.00	0.00
标准差	8.99	5.36	3.53	2.91	2.22
方差	80.87	28.72	12.46	8.47	4.92
范围	115.00	33.00	20.00	18.00	11.00
最小值	0.00	0.00	0.00	0.00	0.00
最大值	115.00	33.00	20.00	18.00	11.00
满足条件的参考文献的数量	248	248	247	217	171

表5.8 参考文献对应的名词性短语个数（50%-100%）

	50	60	70	80	90	100
平均值	0.40	0.33	0.07	0.04	0.01	0.01
中位数	0.00	0.00	0.00	0.00	0.00	0.00
众数	0.00	0.00	0.00	0.00	0.00	0.00
标准差	0.84	0.80	0.26	0.19	0.11	0.09
方差	0.70	0.64	0.07	0.04	0.01	0.01
范围	9.00	9.00	1.00	1.00	1.00	1.00
最小值	0.00	0.00	0.00	0.00	0.00	0.00
最大值	9.00	9.00	1.00	1.00	1.00	1.00
满足条件的参考文献的数量	77	63	19	10	3	2

不同的百分数之下被引次数超过十次的参考文献的每篇参考文献对应的引文上下文当中的名词性短语个数的统计结果可以参见表 5.9 和表 5.10。以被引次数超过十次的参考文献作为研究对象的时候，在百分数为 0 的情况下，可以看

到，平均每篇参考文献里，满足条件的名词性短语有 17.11 个，明显大于以全部参考文献作为研究对象时相同条件下的平均值。这个原因还是因为较多的被引次数就意味着有较多的施引文献，有较多的施引文献就可能有更多的引文句，从而可能有更多的名词性短语。所以，相比于全部参考文献，被引次数超过十次的参考文献在初始阶段的平均值会明显较大一些。然而，在百分数达到三十的时候，即参考文献的名词性短语要出现在该参考文献对应的百分之三十及以上的施引文献当中的情况时，统计结果当中的众数就已经变为 0，这表示，在此时，不满足条件的参考文献的数量已经超过了满足条件的数量。在百分数增加到为四十的情况下，平均值就已经降低为 0.76，而中位数和众数则降低为 0。在百分之七十的情况下，平均值就只有 0.02 了，而满足条件的参考文献则只剩下一个。在百分之七十以上的情况下，平均值就已经为 0，表示被引次数在十次以上的 44 篇参考文献已经没有满足条件的参考文献了。相比于以全部参考文献作为研究对象时的情况下，被引次数超过十次的参考文献在较高百分数的层级上满足条件的参考文献数量较少，主要的原因是较多的施引文献就会带来更多的变化性。对于一篇只有五篇施引文献的参考文献，只要有某一个名词性短语同时出现在这五篇施引文献的引文上下文当中，这篇参考文献就满足了百分之百层级的要求，但是，对于一篇拥有二十篇施引文献的参考文献，只有当其中的某一个名词性短语同时出现在所有二十篇施引文献的引文上下文当中的时候，这篇参考文献才能够满足百分之百层级的要求。更多的施引文献意味着有可能更多的作者对一篇文章进行引用，而每一篇文章的作者都会对一篇参考文献有不同的体会和看法，会从当中吸取到不同的知识，获得不同的概念，这种随施引文献增多而带来的更加丰富的多样性将无疑会使同样表达方式的概念出现在全部或者大部分的施引文献当中的几率减小。所以，相比于被引次数较少的参考文献，较高被引参考文献拥有更多概念及概念表达方式的多样性，从而也使得同一个名词性短语出现在较高比率的施引文献的几率降低。

表5.9 十次以上被引参考文献对应的名词性短语个数（0%-40%）

	0	10	20	30	40
平均值	17.11	13.37	3.76	1.39	0.76
中位数	14.50	13.00	3.00	1.00	0.00
众数	18.00	18.00	3.00	0.00	0.00
标准差	16.68	6.37	2.77	1.58	1.04
方差	278.19	40.64	7.65	2.51	1.07
范围	113.00	31.00	16.00	6.00	5.00
最小值	2.00	2.00	0.00	0.00	0.00
最大值	115.00	33.00	16.00	6.00	5.00

表 5.9 十次以上被引参考文献对应的名词性短语个数（0%-40%）（续）

	0	10	20	30	40
满足条件的参考文献的数量	44	44	43	27	21

表5.10 十次以上被引参考文献对应的名词性短语个数（50%-100%）

	50	60	70	80	90	100
平均值	0.22	0.15	0.02	0.00	0.00	0.00
中位数	0.00	0.00	0.00	0.00	0.00	0.00
众数	0.00	0.00	0.00	0.00	0.00	0.00
标准差	0.47	0.42	0.15	0.00	0.00	0.00
方差	0.22	0.18	0.02	0.00	0.00	0.00
范围	2.00	2.00	1.00	0.00	0.00	0.00
最小值	0.00	0.00	0.00	0.00	0.00	0.00
最大值	2.00	2.00	1.00	0.00	0.00	0.00
满足条件的参考文献的数量	9	6	1	0	0	0

对每一篇参考文献所对应的出现在不同参考文献的施引文献篇数最多的名词性短语，它们出现的施引文献篇数的统计结果如表 5.11 所示，在结果当中除了从被引次数角度而分析被引次数大于 10 的参考文献，还区分了名词性短语的完全匹配，和名词性短语的包含匹配。完全匹配数值的含义较为容易理解，实际上就是某一个名词性短语出现在不同施引文献的个数，例如某参考文献有施引文献五篇对应的一个名词性短语“h index”出现在这篇参考文献的四篇施引文献并且其它名词性短语都出现在三篇及以下的施引文献当中，是出现施引文献篇数最多的名词性短语，这个名词性短语完全匹配的数值就等于四。名词性短语包含匹配的含义与完全匹配不同，指的是只要施引文献当中有名词性短语的字符串当中包含参考文献所对应的出现在不同参考文献的施引文献篇数最多的名词性短语，那么不管这篇施引文献有没有名词性短语与这个名词性短语完全匹配，名词性短语的包含匹配的数值计数中都会包括这篇施引文献。还是上面的例子，“h index”是对应一篇参考文献出现施引文献篇数最多的名词性短语，五篇施引文献的余下那篇施引文献并没有“h index”与之匹配，所以从完全匹配的角度来讲数值是四，但是，这篇施引文献当中却有“web h index”的名词性短语的出现，而这个名词性短语完全可以包含“h index”，那么，从包含匹配的角度来讲，“h index”这个名词性短语的数值就增加到了五，从而全

部覆盖了参考文献全部的施引文献。在统计结果当中可以看到，包含匹配的平均值是 4.56 而完全匹配的平均值是 3.53，同时包含匹配的最大值是 73 而完全匹配的最大值是 61。对于被引次数大于 10 次的参考文献，包含匹配的平均值是 9.54 而完全匹配的平均值是 6.78 都大于全部参考文献时的平均值。

表5.11 名词性短语对应施引文献篇数的统计结果

	A	B	C	D
平均值	3.53	4.56	6.78	9.54
中位数	3	4	5	7.5
众数	3	3	3	6
标准差	4.04	5.10	8.66	10.41
方差	16.33	26.03	75.06	108.30
范围	60	72	59	71
最小值	1	1	2	2
最大值	61	73	61	73

A: 完全匹配（全部参考文献） B: 包含（全部参考文献） C: 完全匹配（被引次数大于 10 的参考文献） D: 包含（被引次数大于 10 的参考文献）

每个参考文献出现在它的施引文献篇数最多的名词性短语，在不同匹配施引文献篇数下的完全匹配和包含匹配的频数和百分率的情况，可以参见表 5.12 和表 5.13。注意到尽管表格里的数据针对的是名词性短语，但是，这些名词性短语都是每篇参考文献当中出现在该参考文献的施引文献篇数最多的名词性短语，因此，这些名词性短语与相应的参考文献有着一一对应的关系，所以，名词性短语总的频数也等于全部参考文献篇数总的篇数。可以从两个表格里看出，无论是包含匹配还是完全匹配，占有最高百分率的匹配施引文献个数都是 3。

表5.12 名词性短语完全匹配的不同施引文献篇数的频数和百分比

匹配施引文献个数	频数	百分率	有效百分率	累积百分率
1	15	5.70	5.70	5.70
2	81	30.80	30.80	36.50
3	87	33.08	33.08	69.58
4	39	14.83	14.83	84.41
5	12	4.56	4.56	88.97
6	8	3.04	3.04	92.02
7	10	3.80	3.80	95.82
8	6	2.28	2.28	98.10

表 5.12 名词性短语完全匹配的不同施引文献篇数的频数和百分比（续）

匹配施引文献个数	频数	百分率	有效百分率	累积百分率
9	2	0.76	0.76	98.86
10	1	0.38	0.38	99.24
19	1	0.38	0.38	99.62
61	1	0.38	0.38	100.00
总计	263	100	100	

表5.13 名词性短语包含匹配的不同施引文献篇数的频数和百分比

匹配施引文献个数	频数	百分率	有效百分率	累积百分率
1	14	5.32	5.32	5.32
2	48	18.25	18.25	23.57
3	67	25.48	25.48	49.05
4	45	17.11	17.11	66.16
5	33	12.55	12.55	78.71
6	14	5.32	5.32	84.03
7	13	4.94	4.94	88.97
8	10	3.80	3.80	92.78
9	3	1.14	1.14	93.92
10	6	2.28	2.28	96.20
11	3	1.14	1.14	97.34
12	1	0.38	0.38	97.72
13	1	0.38	0.38	98.10
14	1	0.38	0.38	98.48
15	1	0.38	0.38	98.86
19	1	0.38	0.38	99.24
23	1	0.38	0.38	99.62
73	1	0.38	0.38	100.00
总计	263	100	100	

完全匹配篇数和包含匹配篇数占施引文献总篇数百分数的描述性统计结果可以参见表5.14。对于全部参考文献，完全匹配的百分比的平均值大约为45.10，包含匹配的百分比的平均值大约为 55.99。对于被引十次以上的参考文献，完全匹配的百分比的平均值大约为 39.09，接近于百分之四十，而在包含匹配的条件下，百分比的平均值大约为 56.15，相比于完全匹配的结果有了较为明显的提升。总体而言，在完全匹配的情况下，较高被引次数的参考文献在百分比平均值反

而不如全部参考文献，可见较低被引次数的参考文献的名词性短语反而会出现在较高比例的施引文献当中。在包含匹配的情况下，全部参考文献百分比略微小于被引次数十次以上的参考文献。

表5.14 完全匹配篇数和包含匹配篇数占施引文献总篇数百分数

	A	B	C	D
平均值	45.10	55.99	39.09	56.15
中位数	40.00	55.56	38.80	54.55
众数	40.00	40.00	21.43	50.00
标准差	15.24	21.28	14.82	20.24
方差	232.16	452.98	219.67	409.54
范围	63.33	85.71	56.64	77.38
最小值	16.67	14.29	14.29	14.29
最大值	80.00	100.00	70.93	91.67

A: 完全匹配（全部参考文献） B: 包含匹配（全部参考文献） C: 完全匹配（被引十次以上的参考文献） D: 包含匹配（被引十次以上的参考文献）

结 论

本研究着眼于引文上下文当中的概念抽取，利用了表示概念的名词性短语可以在施引文献所构成的可比语料反复出现的特点，在对引文上下文分割出名词性短语的情况下，将在引文上下文多次出现的名词性短语提取出来。尽管，前人采取过类似的方法对名词性短语进行提取，但是与前人的研究相比，本研究有以下几点不同。

首先，本研究所构建的系统可以达到自动化或者接近完全自动化。尽管在参考文献信息的匹配的时候，系统在极端的情况下要求人工进行参与，但是这是由于科学引文索引参考文献数据的不完整性所造成的。如果采取 CiteSeer 的复杂方案，那么，科学引文索引参考文献数据的不完整性所造成的人工参与的情况将会完全得到避免。本研究将重点放在引文上下文的概念抽取上，引文数据库的构建只是预处理当中的一个重要步骤，并且已经存在 CiteSeer 一类的成熟但却复杂的方案。在本研究的数据规模下，与构建 CiteSeer 所使用的复杂方法相比，本研究当中所采用的通过科学引文索引数据和 HistCite 软件构建引文数据库的方法是较为合适的。在实验数据下，也并没有出现人工参与的情况。可以讲，本研究的系统达到了自动化或者接近完全自动化的程度。

其次，前人的研究只是针对参考文献当中具有最高被引次数的一类参考文献，在进行概念抽取的过程当中，对于每一个参考文献也只是从这个参考文献众多的施引文献当中随机选取一部分施引文献作为概念选取的语料。在本研究当中，在一个较小阈值之上的全部参考文献都会被挑选出来作为研究对象，并且每一篇参考文献的全部施引文献都会被作为概念抽取的语料。在这种情况下，研究针对的就不仅仅是较高被引文献的情况，也同时包含了被引次数在中等程度和被引次数较低的参考文献，只有被引次数处于阈值之下的参考文献会被排除研究的范围。同时，将参考文献的全部施引文献都作为概念抽取的语料，避免了从全部施引文献当中随机挑选部分施引文献作为语料所可能带来的造成结果不准确的问题。

最后，本研究首次提出了将包含提取出来的名词性短语的名词性短语也在某种程度上作为该名词性短语的扩展，并从实验结果当中发现了，包含形式名词性短语将较为显著的覆盖较多的施引文献，同时，包含形式名词性短语也扩展了概念的表达形式，带来了新的研究问题。

参考文献

[1] LU Z. PubMed and beyond: a survey of web tools for searching biomedical literature [J]. Database: the journal of biological databases and curation, 2011, 3(1): 69-74.

[2] HUNTER L, COHEN K B. Biomedical language processing: perspective what's beyond PubMed? [J]. Molecular cell, 2006, 21(5): 589-594.

[3] GARFIELD E. Science citation index-a new dimension in indexing [J]. Science, 1964, 144(3619): 649-654.

[4] GARFIELD E. The citation index as a subject index [J]. Current contents, 1974, 18(5): 7-9.

[5] GARFIELD E. Citation indexes for science: a new dimension in documentation through association of ideas [J]. Science, 1955, 122(3159): 108-110.

[6] ADAIR W C. Citation indexes for scientific literature? [J]. American Documentation, 1955, 6(1): 31-32.

[7] WEINSTOCK M. Citation indexes [M]. Philadelphia: Institute for Scientific Information, 1971.

[8] NAKOV P I, SCHWARTZ A S, HEARST M. Citances: citation sentences for semantic analysis of bioscience text [C]. Proceedings of the SIGIR'04 workshop on search and discovery in bioinformatics. 2004: 81-88.

[9] WHITE H D. Citation analysis and discourse analysis revisited [J]. Applied Linguistics, 2004, 25(1): 89-116.

[10] MORAVCSIK M J, MURUGESAN P. Some results on the function and quality of citations [J]. Social studies of science, 1975, 5(1): 86-92.

[11] GARFIELD E. Can citation indexing be automated [C]. Statistical Association Methods for Mechanized Documentation. 1965: 189-192.

[12] THORNE F C. The citation index: another case of spurious validity [J]. Journal of Clinical Psychology, 1977, 33(4): 1157-1161.

[13] O'CONNOR J. Citing statements: computer recognition and use to improve retrieval [J]. Information Processing & Management, 1982, 18(3): 125-31.

[14] O'CONNOR J. Biomedical citing statements: computer recognition and use to aid full-text retrieval [J]. Information Processing & Management, 1983, 19(6): 361-8.

[15] SMALL H G. Cited documents as concept symbols [J]. Social studies of science, 1978, 8(3): 327-340.

[16] LAWRENCE S, LEE GILES C, BOLLACKER K. Digital libraries and autonomous citation indexing [J]. Computer, 1999, 32(6): 67-71.

[17] BRADSHAW S G, ADVISER-HAMMOND K. Reference directed indexing:

indexing scientific literature in the context of its use [M]. Northwestern University, 2002.
[18] BRADSHAW S. Reference directed indexing: Redeeming relevance for subject search in citation indexes [J]. Research and Advanced Technology for Digital Libraries, 2003, 499-510.
[19] RITCHIE A. Citation context analysis for information retrieval [D]. Cambridge: University of Cambridge, 2008.
[20] ALJABER B, STOKES N, BAILEY J, et al. Document clustering of scientific texts using citation contexts [J]. Information retrieval, 2010, 13(2): 101-131.
[21] 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.
[22] NANBA H, OKUMURA M. Towards multi-paper summarization reference information [C]. IJCAI'99 Proceedings of the 16th international joint conference on Artificial intelligence - Volume 2. Morgan Kaufmann Publishers Inc, 1999: 926-931.
[23] NANBA H, KANDO N, OKUMURA M. Classification of Research Papers using Citation Links and Citation Types: Towards Automatic Review Article Generation [C]. 11th ASIS SIG/CR Classification Research Workshop. 2000: 117-136.
[24] NANBA H, ABEKAWA T, OKUMURA M, et al. Bilingual presri: Integration of multiple research paper databases [C]. Proceedings of the conference on large scale semantic access to content (RIAO). 2004: 195-211.
[25] NANBA H, OKUMURA M. Automatic detection of survey articles [J]. Research and Advanced Technology for Digital Libraries, 2005, 3642: 391-401.
[26] TEUFEL S, MOENS M. Summarizing scientific articles: experiments with relevance and rhetorical status [J]. Computational linguistics, 2002, 28(4): 409-45.
[27] TEUFEL S, SIDDHARTHAN A, TIDHAR D. Automatic classification of citation function [C]. EMNLP '06 Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2006: 103-110.
[28] ELKISS A, SHEN S, FADER A, et al. Blind men and elephants: what do citation summaries tell us about a research article? [J]. Journal of the American Society for Information Science and Technology, 2008, 59(1): 51-62.
[29] MOHAMMAD S, DORR B, EGAN M, et al. Using citations to generate surveys of scientific paradigms [C]. The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics. NAACL '09 Proceedings of Human Language Technologies. Association for Computational Linguistics, 2009: 584-592.
[30] W. SCHNEIDER J. Concept symbols revisited: naming clusters by parsing and

filtering of noun phrases from citation contexts of concept symbols [J]. Scientometrics, 2006, 68(3): 573-593.
[31] GARFIELD E, PARIS S, STOCK W G. HistCite™: a software tool for informetric analysis of citation linkage [J]. Information Wissenschaft und Praxis, 2006, 57(8): 391.
[32] GARFIELD E, PUDOVKIN A, ISTOMIN V. Algorithmic citation-linked historiography—mapping the literature of science [J]. Proceedings of the American Society for Information Science and Technology, 2002, 39(1): 14-24.
[33] CROFT B, METZLER D, STROHMAN T. Search engines: information retrieval in practice [M]. Addison Wesley, 2010.
[34] MARCUS M P, MARCINKIEWICZ M A, SANTORINI B. Building a large annotated corpus of English: The Penn Treebank [J]. Computational linguistics, 1993, 19(2): 313-330.
[35] JURAFSKY D, MARTIN J H. Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition [M]. Prentice Hall, 2000.
[36] BIRD S, KLEIN E, LOPER E. Natural language processing with Python [M]. O'Reilly Media, 2009.

附 录

实验当中所使用的263参考文献的编号、参考文献对应施引文献数量、参考文献对应引文上下文出现最多次数的名词性短语、以及该名词性短语完全匹配和包含匹配的施引文献数量如下所示：

编号	施引文献数量	名词性短语	完全匹配	包含匹配
562	5	CC	2	3
613	6	citation	2	3
715	12	SDSs	5	5
716	11	application	3	3
719	6	Italian system	2	2
741	6	Scientific Information	1	1
770	6	highly	3	3
771	5	citation	4	5
772	6	positive correlation	3	3
804	8	h index	6	7
833	12	h index	8	10
845	5	English language journal	2	2
978	6	GS	2	3
995	8	productivity	3	3
1000	5	field	2	3
1059	5	citation	2	2
1179	8	network	3	5
1235	10	co citation	4	5
1251	6	citation	3	4
1258	6	NMCR	3	4
1407	5	author	3	3
1418	6	improvement	3	3
1424	12	Google Scholar	6	7
1428	13	network	3	10
1430	16	collaboration	4	11
1458	5	study	3	4
1467	16	paper	8	8
1476	6	equipment	2	2

1563	7	citation	2	4
1564	6	citation	2	3
1569	6	university	4	4
1593	11	field	3	3
1600	7	h index	4	5
1604	10	h index	7	7
1606	5	manuscript	3	3
1607	9	h index	7	7
1608	10	citation	3	8
1648	12	science	5	5
1663	8	nanotechnology	3	4
1673	13	h index	7	8
1692	8	Google	3	5
1724	6	h	4	6
1725	5	Another tangible sign	2	2
1738	7	effect	3	5
1739	5	CWTS	2	2
1745	10	Co word analysis	3	5
1748	7	network	3	5
1838	5	DEA	2	5
1942	6	absorptive capacity	2	2
1952	6	author	3	3
2034	6	work	6	6
2047	5	author	2	2
2089	5	assumption	2	2
2176	5	Domain visualization	2	2
2182	11	CiteSpace	5	6
2186	5	brokerage process	2	2
2188	5	Taiwan	2	2
2287	9	h index	4	5
2305	5	book	2	2
2306	6	hyperauthorship	3	4
2310	7	scientist	3	4
2317	6	country	5	6
2375	6	Pajek	2	3
2632	5	co word analysis	3	4
2697	11	source	5	6

2792	9	h index	4	5
2793	28	g index	19	23
2794	5	h index	5	5
2797	6	h index	4	4
2805	8	h index	4	4
2839	10	university	7	7
2959	5	fact	3	3
2981	8	betweenness centrality	3	4
2982	17	network	8	13
3114	6	attention	2	2
3175	8	Impact Factor IF	2	3
3190	15	journal	6	10
3199	14	Citations	2	2
3283	12	science	6	6
3335	7	collaboration	4	7
3354	6	C2	4	4
3470	8	patent	4	6
3567	10	journal	6	8
3596	9	advantage	2	2
3597	5	researcher	3	3
3610	5	author	2	2
3619	6	China	2	2
3620	7	international collaboration	3	3
3621	15	international collaboration	4	8
3623	12	JIF	3	3
3626	11	field	3	7
3636	5	author	3	3
3637	11	h index	5	7
3638	6	country	3	3
3642	5	China	2	2
3644	5	indicator	2	3
3699	5	job	2	2
3725	5	China	4	4
3769	5	research	2	5
3807	5	group	2	3
3819	6	patent value	4	4
3892	6	university	2	4

4061	5	authorship credit	3	3
4072	12	patent citation	6	7
4176	86	h index	61	73
4177	11	h index	7	7
4210	5	Scientometrics	3	4
4242	8	study	3	3
4297	5	South African scientist	3	3
4321	6	firm	4	4
4324	6	patent citation	3	4
4367	7	A index	4	6
4420	6	source	3	3
4451	5	South Africa	3	3
4468	15	h index	9	10
4483	6	collaboration	2	2
4511	5	Coupe	1	1
4514	7	Paper	2	2
4530	6	linkage	2	2
4543	21	collaboration	10	19
4703	10	contribution	2	2
4704	5	science	2	4
4775	6	h index	3	3
4806	22	country	8	15
4818	6	science	3	5
4863	5	China	2	2
4883	5	citation	2	3
4885	6	GS	2	2
5097	5	Scientometrics	3	3
5169	8	Shanghai ranking	3	3
5189	5	woman	4	4
5204	13	author	5	6
5243	8	researcher	3	4
5244	7	Absolute	1	1
5256	5	co authorship	2	3
5282	5	WoS	2	2
5298	12	collaboration	7	11
5328	7	age	3	3
5334	8	Patent citation	3	3

5347	5	ecology	2	3
5367	5	1990s	1	1
5403	6	journal	3	5
5404	5	www. leydesdorff. net	1	1
5409	6	China	3	3
5417	7	indicator	3	3
5420	8	nanotechnology	3	5
5425	5	field	2	3
5431	6	correlation	2	2
5433	10	science	7	8
5436	5	U. S .	2	2
5453	5	field	2	2
5489	6	conference proceeding	4	4
5502	6	Data SIGMOD	2	2
5536	8	citation	4	8
5564	7	citation	3	4
5687	5	author	3	3
5688	5	4Cij	1	1
5749	15	science	4	6
5754	5	11For example many sociologist	1	1
5813	23	publication	9	12
5815	5	peak year citation	2	2
5825	14	CPPm	3	5
5943	5	citation	2	5
5949	8	citation	2	5
6021	8	country	4	4
6042	5	homophily	3	3
6048	19	Google Scholar	6	8
6052	13	collaboration	5	11
6071	8	nanotechnology	4	5
6072	7	nanotechnology	3	3
6073	5	nanotechnology	3	3
6074	6	patent	3	5
6081	6	co activity	2	2
6084	7	field	3	5
6086	6	field	3	3
6119	7	UK	2	3

6121	5	journal	4	5
6138	6	collaboration	3	4
6142	5	interdisciplinarity	3	3
6194	5	science	5	5
6209	9	patent	3	9
6211	9	technology	3	6
6327	6	research	2	2
6337	5	humanity	3	4
6356	13	collaboration	4	6
6360	5	network	2	4
6362	10	author	4	4
6387	5	coverage	3	3
6395	5	Adverse drug reaction	1	2
6428	5	British research	2	2
6477	5	Center	1	1
6492	7	bibliometrics	3	3
6589	5	Other map	1	1
6615	8	citing journal	4	4
6658	5	stemming algorithm	2	4
6692	6	demographic	2	2
6693	16	field	4	4
6704	14	citation	4	8
6706	5	Journal	1	1
6711	7	1960s	1	1
6719	5	statistical method	3	3
6774	5	Paper	2	2
6792	10	co authorship	3	3
6803	6	citation	2	3
6842	6	Interdisciplinarity	2	4
6843	8	nanotechnology	3	5
6845	5	field	2	3
6847	5	interdisciplinarity	4	4
6860	6	Africa	4	4
6885	5	woman	3	3
6995	9	positive correlation	3	3
6996	7	biotechnology	3	3
7105	5	citation	2	4

7146	6	power law distribution	3	3
7240	5	h index	4	4
7374	12	journal	3	4
7377	5	journal	2	2
7378	5	international collaboration	3	4
7414	5	distribution	2	3
7521	14	document	3	5
7526	7	co citation	2	3
7576	5	company	2	2
7665	8	collaboration	4	7
7693	7	attention	2	2
7731	5	men	2	2
7752	6	coauthored publication	2	2
7763	5	scientist	2	2
7766	5	inverse	3	3
7774	18	h index	8	9
7777	5	h index	3	5
7783	9	field	4	4
7848	11	h index	7	8
7880	6	patent	3	3
7884	5	Paper	1	1
7923	5	SCI journal	2	2
8030	5	patent	3	4
8103	10	patent	7	9
8204	5	Africa	3	3
8412	5	First	1	1
8415	6	journal	2	3
8423	7	peer review	2	2
8424	5	international collaboration	3	3
8425	14	author	3	3
8426	19	h index	8	10
8432	7	h index	5	6
8557	17	network	4	14
8606	6	author	2	5
8616	5	mutual dependency	2	2
8694	11	collaboration	3	10
8731	5	citation	2	4

8738	9	Watts	5	5
8762	10	ACA	2	2
8766	7	Pathfinder	2	6
8795	10	h index	4	4
8854	5	network	3	5
8905	6	China	4	4
9004	6	e index	4	4
9034	9	China	7	7
9058	6	journal	3	5
9085	5	h index	3	3
9094	7	HCP	1	1
9095	14	collaboration	3	7
9096	8	Medline	2	3

作者简介

姓名：孙枫军
性别：男
民族：汉族
出生年月：1987-08
籍贯：山东

教育经历：

2006-09—2010-07	东北大学	软件工程专业	学士
2010-09—2013-04	中国科学技术信息研究所	情报学专业	硕士

参加项目：

国家“十二五”科技支撑计划项目《面向外文科技文献信息的知识组织体系建设与应用示范》中的子课题：《面向外文科技知识组织体系的大规模语义计算关键技术研究》(项目编号：2011BAH10B04).

攻读硕士学位期间发表的学术成果

论文

[1] Sun F, Zhu L. Citation genetic genealogy: a novel insight for citation analysis in scientific literature [J]. Scientometrics, 2012, 91(2): 577-58

[2] Sun F. Citation genetic genealogy: a new perspective for citation analysis in scientific literature [C]. Proceedings of the ISSI 2011 Internatinal Conference. 2011: 817-828.

[3] Sun F., Zhu L. Using association rule to find one discipline's papers published on another discipline's journals [C]. Proceedings of the AST, EEC, MMHS, AIA 2012 International Conference. 2012: 184-189.

软件著作权

[1] 引文上下文的概念抽取系统（登记号：2012SR089302）

学位论文数据集

<table>
<tr><td>关键词*</td><td>密级*</td><td>中图分类号*</td><td>UDC</td><td>论文资助</td></tr>
<tr><td>引文上下文、概念抽取、引文数据库、引文索引</td><td>公开</td><td>TP391</td><td></td><td></td></tr>
<tr><td>学位授予单位名称*</td><td colspan="2">学位授予单位代码*</td><td>学位类别*</td><td>学位级别*</td></tr>
<tr><td>中国科学技术信息研究所</td><td colspan="2">80901</td><td>管理学</td><td>硕士</td></tr>
<tr><td>论文题名*</td><td colspan="3">并列题名*</td><td>论文语种*</td></tr>
<tr><td>引文上下文中的概念抽取</td><td colspan="3">无</td><td>中文</td></tr>
<tr><td>作者姓名*</td><td>孙枫军</td><td colspan="2">学号*</td><td>809011013</td></tr>
<tr><td>培养单位名称*</td><td>培养单位代码*</td><td colspan="2">培养单位地址</td><td>邮编</td></tr>
<tr><td>中国科学技术信息研究所</td><td>80901</td><td colspan="2">北京市复兴路 15 号</td><td>100038</td></tr>
<tr><td>学科专业*</td><td>研究方向*</td><td colspan="2">学制*</td><td>学位授予年*</td></tr>
<tr><td>情报学</td><td>知识工程</td><td colspan="2">2.5 年</td><td>2013 年</td></tr>
<tr><td>论文提交日期*</td><td colspan="4">2012 年 10 月</td></tr>
<tr><td>导师姓名*</td><td>朱礼军</td><td>职称*</td><td colspan="2">副研究员</td></tr>
<tr><td>评阅人</td><td>答辩委员会主席*</td><td colspan="3">答辩委员会成员</td></tr>
<tr><td></td><td>赖茂生</td><td colspan="3">靖培栋、王惠临、姚长青、刘耀</td></tr>
<tr><td colspan="5">电子版论文提交格式　文本(√)　图像()　视频()　音频()　多媒体()　其它()
推荐格式：application/msword；application/pdf</td></tr>
<tr><td>电子版论文出版(发布)者</td><td colspan="2">电子版论文出版(发布)地</td><td colspan="2">权限声明</td></tr>
<tr><td></td><td colspan="2"></td><td colspan="2"></td></tr>
<tr><td>论文总页数*</td><td colspan="4">59</td></tr>
<tr><td colspan="5">共 33 项，其中带*为必填数据，为 22 项。</td></tr>
</table>

引文上下文中的概念抽取

作者： 孙枫军
学位授予单位： 中国科学技术信息研究所