项目结题验收单

专家验收表（主持人所在单位组织 3-5 名专家对项目进行验收、自评。）

项目名称	基于近似模式匹配的严格自引检测研究		
主持人	宫兆阳	职务/职称	馆员
所在单位	中国海洋大学图书馆		
专家意见	本项目进行了国内外在自引检测方面的文献分析，归纳并阐述了 3 个研究方向；根据目前人工查收查引过程中在排除自引方面积累的精细化检索经验，分析严格自引检测的相关特征；研究严格自引检测效果评价，主要从准确率(Precision)、召回率(Recall)、F 指标(F-Measure)、时间效率四个维度进行评价研究。 本项目为图书馆在严格自引检测方面的研究，主要为解决传统严格自引检测精确率、召回率、效率低的问题。通过研究国内外的文献，分析严格自引检测的相关特征，严格自引检测效果评价等多个方面进行研究综述，提出近似模式匹配算法应用于严格自引检测以及未来的研究方向，为去除科研文献评价中的自引干扰，增强评价准确度提供了新的解决思路。 基于以上研究，形成《严格自引检测研究综述》调研报告 经专家评议，一致同意通过结题验收。		
专家签字			
职务/职称			

项目编号：2022072
注:项目编号请查看立项通知，也可缺省

CALIS 全国农学文献信息中心研究项目
结题报告

项目名称：基于近似模式匹配的严格自引检测研究

项目关键词：近似模式匹配；严格自引；查收查引

项目单位（盖章）：中国海洋大学图书馆

通信地址：山东省青岛市崂山区松岭路 238 号中国海洋大学图书馆（266100）

项目主持人：宫兆阳

联系电话：15610563026

电子邮件：gongzhaoyang@ouc.edu.cn

提交日期：2023 年 5 月 23 日

基于近似模式匹配的严格自引检测研究

关键词： 近似模式匹配；严格自引；查收查引

1 研究背景、目的及意义

1.1 背景简介

自我引用（以下简称“自引”）是文献计量学和科研评价中的一项非常重要的指标概念，根据不同规模的自引分析，其数值的大小直接影响作者、期刊、科研单位甚至国家的科研水平和实力。国外学者 Shehatta[1]利用 Scopus 数据库分析自引对 1996-2015 年期间总引文量、每篇论文平均引文量、被引出版物百分比以及十个最具研究生产力国家的排名的影响，我国以 54.9%的自引率排名第一，美国以 45.6%的自引率排名第二。研究表明，自引对顶级国家的学术表现有很大影响，因此，需要计算各种基于引文的文献计量学指标来检测排除自引，以期消除全部影响，排除自引干扰。近年来，自引检测相关的研究呈现更加智能化、自动化、精确化的发展趋势，通过分析国内外在文章自引方面的文献，其研究的方向主要包括以下 3 个方面：、结合实践经验，深入研究自引检测方法、查收查引自动化软件设计实施包含自引检测相关内容，该类文献一般只做简略介绍、其他国外的自引研究。

1.2 目的和意义

本课题以研发收录引证自动生成系统为契机，针对其关键处理环节——自引检测，结合近似模式匹配 DP 算法，解决传统严格自引检测精确率、召回率、效率低的问题。课题将根据当下人工查收查引过程中的排除自引积累的精细化的检索经验，分析严格自引检测的相关特征；通过分析传统的精确字符串匹配在严格自引检测中的弊端，揭示其检测的提升空间；设计实验验证课题提出的基于近似模式匹配的严格自引检测，将从准确率(Precision)、召回率(Recall)、F 值(F-Measure)、时间效率四个维度评价该项研究。本课题所提出的研究算法将有效提升文献级严格自引检测的效率，可以实现接近完全的召回率，而只产生可忽略的精确度损失，为去除科研文献评价中的自引干扰，增强评价的准确度上提供了新的解决思路。

2 课题研究方法及思路

本课题将以中国海洋大学图书馆科技查新工作站在查收查引的实际工作为重要的经验基础，结合查收查引自动化系统的设计开发的需求，总结出文献级自引检测是查收查引工作的一项关键的处理任务，其检测效率影响了整个工作站的业务处理能力，其准确性决定了查收查引报告的可靠程度。所以，正确检测和分析作者自引在研究人员的评价性文献计量评估中至关重要。课题定义“严格自引”的概念，以区别于只检测首作者的自引，即：如果被引文献和施引文献作者名单的集合不相交时，换言之，文献之间至少有一个共同的作者，就为“严格自引”。

课题研究将基于近似模式匹配DP（动态规划）算法在字符串模糊匹配方面的突出优势，对两位作者姓名偏差以及对这种偏差的方向和程度进行量化。主要功能为检测出两种相异字符串，即一种因为姓名的同义性而标记两个姓名为一个作者的“假阳性”字符串，以及另一种因为作者姓名出现变体（例如：Zhang San、Zhang S）而错误地将它们分开的“假阴性”字符串。课题将利用近似模糊匹配算法，得到编辑距离模型，从施引文献作者名单中选择高度相似的姓名来增加相关姓名的召回

率。课题根据收录引证自动生成系统的实现机理，针对其关键处理环节——自引检测，展开的国内外文献相关调查研究，解决传统严格自引检测精确率、召回率、效率低的问题。主要的研究方法及思路如下：

（1）、首先，根据参考文献，对课题背景做深入调研，同时，利用当下人工查收查引过程中的排除自引积累的精细化的检索经验，分析严格自引检测的相关特征，分析国内外在自引检测方面的文献，归纳并阐述了 3 个研究方向；

（2）、其次，分析传统的精确字符串匹配在严格自引检测中的弊端，根据检测结果，总结其提升空间以及优化方案；

（3）、再次，深入研究基于近似模式匹配算法，设计实验验证本课题提出的基于近似模式匹配的严格自引检测，主要从准确率(Precision)、召回率(Recall)、F 值(F-Measure)、时间效率四个维度评价该项研究；

（4）、最后，总结本课题的研究结论并提出了算法应用的局限性与未来的研究方向。

3 课题研究的具体内容

3.1 自引检测的国内外相关研究

自引作为影响科研成果评价的一项重要文献计量指标，是科研工作者、相关单位和研究领域共同关注的一项数据表现。近年来，自引检测相关的研究呈现更加智能化、自动化、精确化的发展趋势。通过分析国内外在文章自引方面的文献，其研究的方向主要包括以下 3 个方面：

（1）结合实践经验，深入研究自引检测方法。尹相权等[2]通过作者姓名、所在机构、学科以及刊源作为规则要素，设计一套人名消歧算法，主要解决中文重名引起的自引检测不够精确的问题，并未涉及外文自引检测的研究；任珩等[3]结合查收查新实践经验，利用python语言设计一套他引自引检测判别软件，主要介绍了软件的设计实现，并未介绍排除自引的算法或者研究方法。周文云等[4]以美国控制学专家Lotfi Zadeh的提出的模糊理论为基础，通过对作者姓名、地址、单位等元素片词模糊，解决传统精确匹配漏检、误检的问题，实现查收查引过程中排除自引的需求，该方案设计与本文提出的近似模式匹配算法比较类似，但属于两种不同的研究方向。国外学者Donner等[5]研究利用模糊作者名匹配和互补误差估计来增强自引检测，将模糊字符串匹配算法用于自引用检测，并将该方法与其他基于专有作者姓名的常见自引用检测方法对人工抽取的Ground Truth样本进行了基准测试。

（2）因为排除自引是查收查引自动化系统的核心关键步骤，查收查引自动化软件设计实施包含自引检测相关内容，该类文献一般只做简略介绍，主要介绍自动化软件的功能实现。马海收等[6]基于ISI Web of Knowledge 平台设计查收查引报告生成软件，以忽略作者姓名中的特殊字符和字母大小写为研究思路，优先解决自引排除召回率低的问题。马芳珍[7]等通过对calis和北京大学图书馆联合开发的收录引证系统进行评价，总结系统包括他引自引检测功能，仍然需要大量的人工干预来应对难以区分的个别问题。王学勤等[8]为了解决中科院文献情报中心开发的引用统计工具不能在线自动化操作的问题，把自引排除作为一项重要需求来研究，但并没有介绍具体的自引排除实施过程。刘彦君等[9]发明公开了一种基于引文网络与科研合作网络的领域专家遴选的方法。唐向东等[10]本发明提供一种申请人自我引用分

析系统，均偏向于系统介绍，并未涉及实际算法研究。

以上的各类系统都是对查收查引工作自动化实现做出的科学专业规范的探索，但是对于严格自引检测方面都没有做更为细致的深入研究。

（3）其他国外的自引研究。Tsay 等[11]从 SSCI journal Citation Reports on Web 2005 对经济学、心理学和政治学期刊的自引率(包括自引率和自引率)和其他科学计量学数据进行了分析和比较。Hartley 等[12]研究作者的自我引用和影响因素，作者自我引用是影响期刊影响力的另一个因素。Ioannidis 等[13]讨论了不同类型的自引的含义,包括直接自引、合著自引、合作自引和强制诱导自引等自引的概括性观点，介绍了一些极端自引实践的案例研究。Aksnes 等[14]分析了超过 4.5 万份出版物，研究调查了自引在挪威科学生产中的作用(1981-1996 年），对自引的宏观进行研究。Thijs 等[15]通过欧洲大学的例子，回答作者自引对文献计量介观指标的影响在多大程度上偏离了宏观层面的影响，以及在文献计量分析中可以在多大程度上使用国家参考标准。Glanzel 等[16]分析了作者自引在文献传播过程中的作用，旨在发现自引在文献传播过程中的基本规律，从而为任何层次的实证研究中对自引模式的批判奠定方法论基础。

综上所述，分析以上对严格自引检测的研究，发现其相关内容是广大学者探讨的热点，但是均未深入结合相关算法进行研究，还是一个非常值得探索的领域。

3.2 严格自引检测的特征

根据以往人工查收查引过程中排除自引积累的精细化的检索方法及流程经验，对严格自引的特征进行分析，得出以下几个方面的特征：

（1）委托量的逐年递增，凸显自引检测的重要性。以中国海洋大学图书馆参考咨询部每年接收的委托单和检索文献数量为例，大概为 2000 个/年的委托申请，检索文章数量大概为 40000 篇，而实际委托人数和文献检索数量需求要大于这个数量，因为工作人员有限，工作时间有限，大量的检索需求和繁杂的检索步骤，明显与现如今的人力资源不相适应，所以查收查引自动化系统的实施是势在必行的，而其中自引检测更是系统的核心功能。

（2）作者数量直接影响严格自引的统计基数。根据经验数据（学校海洋科技名词收集的 136 万论文题录信息）统计出来的作者姓名平均数为 5 个，学校机构知识库中，姓名变体数量一般为 20 个，假设文章的施引文献有 1000 个，那么一篇被检索文章的匹配计算量大概为 10 万，下图为文献中姓名数量的频率统计。

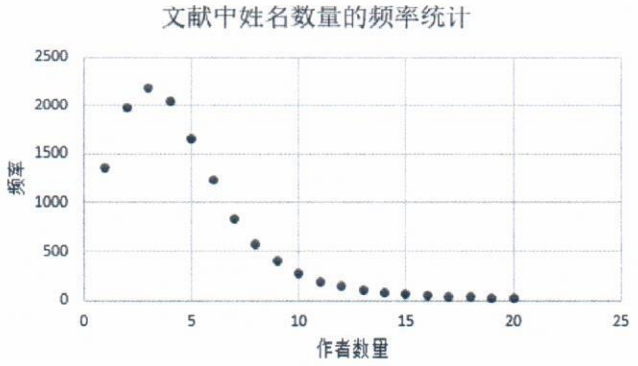

图 1 文献中姓名数量的频率统计

（3）作者姓名变体繁杂加大严格自引检测难度。由于英文姓名在表达上存在多种变体，自引检测的关键点就是消除这种干扰信息，例如：张三，英文的拼音表达

形式可以为San Zhang、Zhang San、ZhangS等。如果姓名字数在2个以上，将会增加更多的姓名拼音表达的可能性。据统计一个三字的姓名，大概变体数在30个左右,严格自引检测的主要功能为检测出两种相异字符串，即一种因为姓名的同义性而标记两个姓名为一个作者的“假阳性”字符串，以及另一种因为作者姓名出现变体（例如：Zhang San、Zhang S）而错误地将它们分开的“假阴性”字符串,“假阳性“和”假阴性“的姓名判断主要依据变体种类来区分，一篇复杂文章的严格自引，或者多篇文章一起来做查收查引，需要多个人配合完成，就需要大量人工时间完成收录引证的工作。

3.3 严格自引检测的评价

在统计学和文献信息检索研究领域，最常用的评价指标为“P&R”值的计算，即准确率与召回率（Precision & Recall）[17]。其中准确率是衡量统计分析和信息检索的正确数据占比，是针对检测结果而言的一项指标，检测为正就有两种可能，一种就是把正类检测为正类(TP)，另一种就是把负类检测为正类(“假阳性”FP)，从统计学的角度解释是检测正类的结果中真实为正的结果数量比例；召回率是衡量统计分析和信息检索的查全率，是针对检测样本而言的一项指标，同样有两种可能性发生，一种是把原来的正类检测成正类(TP)，另一种就是把原来的正类检测为负类(“假阴性”FN)，从统计学的角度解释是被检测出来的正类样本占真实全部正类样本的比例。准确率与召回率的计算公式如图2所示，统计分析关系如图3所示。

$$P = \frac{TP}{TP + FP} \qquad R = \frac{TP}{TP + FN}$$

图2 准确率与召回率的计算公式

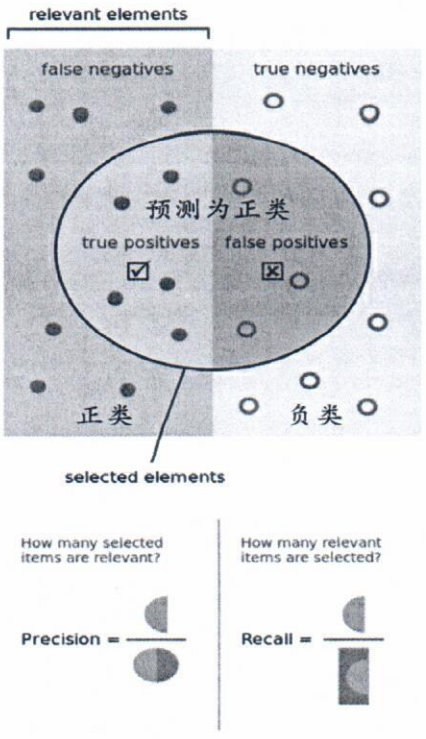

图3 准确率与召回率的统计关系图

以上为基本理论的研究，实际检测分析时，PR指标很多时候表现为互相矛盾的现

象，例如：假设检测结果为 1 个正确结果时，P 指标即为 100%，R 指标就表现为很低；假设检测结果完全等于样本数，R 指标即为 100%，P 指标就表现为很低。所以，PR 指标不能完全表达检测结果时，需要一项更为综合的 F 指标（F-Measure 或 F-Score）来表达。

综合评价指标（F-Measure）通过对准确率和召回率进行加权调和平均，平衡准确率和召回率的不利影响[18]。其中，最常见的为F1综合评价计算方法，如下图所示。

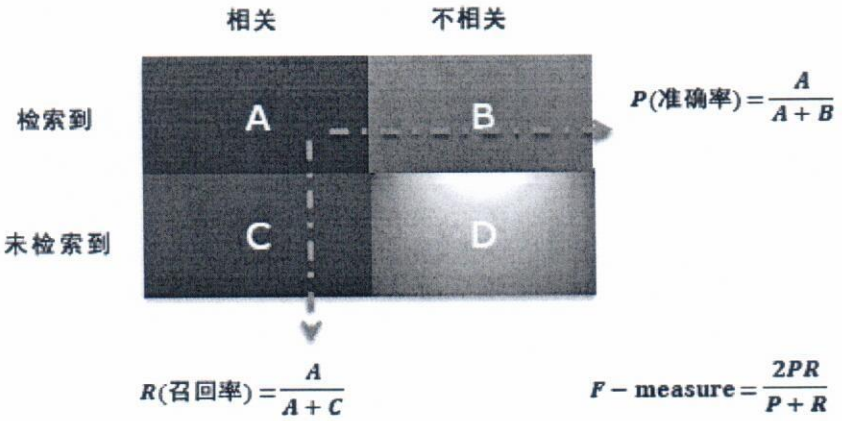

图 4 F1-Measure 综合评价计算方法

在评价严格自引检测的效果时，除了需要 P、R、F 指标的计算，还需要考虑时间效率的指标，P、R、F 指标均为越接近 1 效果越好，时间效率为越低越高效。

3.4 近似模式匹配算法

目前，近似模式匹配算法在检测基因序列、病毒等生物计算领域和语音识别、文件检索、模式识别、OCR、大流量网络信息检测等信息技术领域都得到了广泛的应用，也是算法设计中的一个研究热点[19]。近似模式匹配是相对于精确模式匹配的一种字符串匹配算法，传统精确字符串匹配在严格自引检测中的弊端是显而易见的，其不允许对模式串有任何变化，严格与文本串对比匹配，并不会有效提高检测精确率和召回率。因此文章提出近似模糊匹配算法应用于严格自引检测将会增加更多有益的可能性。

近似模式匹配，在有些文献中也被称为“支持微差异的字符串匹配”，就是按照预先设定的字符串近似相等标准，来找到所有与文本串“相等”的模式串。
其中，设定近似相等差异度的模型为 Levenshtein 距离模型，也成为编辑距离模型[20]。可以理解为：如果对字符串 x 进行 k 个字符的增、删、改操作，可以得到 y 字符，那么认定 x 与 y 是近似相等，使用编辑距离 ed(x,y) 表示字符串 x 与 y 互相转化的最小编辑距离，其中，ed(x,y) <=k。如果字符串长度为 m，那么比值 r=k/m 就称为容错率，通常情况下，我们会根据实验，设计最合理的容错率，对 x、y 的偏差以及对这种偏差的方向和程度进行量化，来保证结果的精确度、召回率和 F 指标最优。

3.5 基于近似模式匹配的严格自引检测实验设计

根据以上理论研究，我们可以设计实验如下：将被引文献的所有作者通过姓名变体生成模式串 P 集合，选择来源于 WOS 检索出来的施引文献作为基础数据源。根据施引文献数量级，随机选择 100、300、500、700、1000、1500、2000 不同级别的

数据作为多组计算数据源，并通过人为干预生成待匹配的文本串 T 样本。首先对模式串 P、文本串 T 进行格式化处理，变成标准的字符串形式，例如处理多余空格、逗号、分号、转大写等字符预处理工作；其次通过人工严格自引检查，确认 100%准确率、100%召回率的自引检测结果，用于后期的数据比对；最后利用 Python 设计近似模式匹配程序处理以上准备数据，得到实验结果。

在结果分析和评价中，可以将三种严格自引检测的结果综合对比，分别为人工自引检测、精确字符串匹配自引检测、近似模式匹配自引检测，结果将从每一篇文献的施引数量、自引数量、他引数量等方面进行数据统计，从精确率、召回率、F指标、时间效率四个维度对比分析，找出其中最有优势的严格自引检测方案。理论分析人工自引检测的准确率、召回率最高，但是人工成本也很高，花费的时间也是最多，即利用无限的时间，可以将结果的精确率和召回率达到100%，这与系统的自动化发展需求不相适应的。精确字符串匹配检测方式，因为不够灵活，将无法检测出各类姓名变体数据，属于精确率、召回率均低的一种模式，但时间效率肯定要比人工检测有优势；最后一种近似模式匹配检测，整体的精确率、召回率都会很高，但是需要持续优化编辑距离模型，否则无法达到接近100%的结果输出，同样其时间效率会比人工检测更有优势。

3.6 未来工作

根据上文的实验设计与结果分析，为进一步发挥近似模式匹配算法在字符串模糊匹配方面的优势，形成应用于严格自引检测方面的高效算法模型，仍然存在以下几个方面的局限性，需要今后做综合考虑和深入研究：

（1）引入其他纠正偏差的参数。为了提高严格自引检测的精确率和召回率，可以引入其他纠正偏差的参数，例如 Author_id，作者单位，地址等信息，在保证时间效率不大幅度增加的情况下，通过多个参数的综合匹配校验，增强检测效果。

（2）错引文献研究。由于出版商或者文献标注引起的错误引用信息，也是很重要的一项研究工作，可以进一步提升文献的实际价值，精确地评价科研论文。

（3）本文设计的近似模式匹配算法，属于单模式匹配算法，其运行效率瓶颈比较明显，未来的研究中，可以针对多模式模糊匹配算法进行研究，实现多线程完成严格自引检测工作。调研关于多模式模糊匹配经典算法（WM算法）的文献，有人通过模糊关键词，再进行多模式精确匹配来变通地实现多模式模糊匹配的功能[21]，这种方式并不能满足严格自引检测的需求。所以，该方向还有很大的研究前景。

4 结论与建议

本课题全面阐述了严格自引检测方面的研究，以研发收录引证自动生成系统为契机，针对其关键处理环节——自引检测，结合近似模式匹配 DP 算法，解决传统严格自引检测精确率、召回率、效率低的问题。首先，根据当下人工查收查引过程中的排除自引积累的精细化的检索经验，分析严格自引检测的相关特征；其次，分析传统的精确字符串匹配在严格自引检测中的弊端；再次，设计实验验证本课题提出的基于近似模式匹配的严格自引检测，主要从准确率(Precision)、召回率(Recall)、F 值(F-Measure)、时间效率四个维度评价该项研究；最后，总结研究结论并提出了算法应用的局限性与未来的研究方向。本课题所提出的研究算法将有效提升文献级严格自引检测的效率，可以实现接近完全的召回率，而只产生可忽略的精确度损失，为去除科研文献评价中的自引干扰，增强评价的准确度上提供了新的解决思路。

5 项目成果

调研报告1篇：《严格自引检测研究综述》

6 参考文献

[1]. Ibrahim, S. and M.A. Abdullah, Impact of country self-citations on bibliometric indicators and ranking of most productive countries. Scientometrics, 2019. 120(2).

[2]. 尹相权，曾姗与糜凯，基于人名消歧的自引统计研究. 情报探索, 2015(05): 第57-59+67页.

[3]. 任珩，基于蟒蛇语言(Python)的查引工作中他自引区分软件的设计与实现. 科技促进发展, 2019. 15(07): 第717-723页.

[4]. 周文云等，基于片词模糊匹配的智能化查收查引系统研究. 情报探索, 2020(10): 第36-41页.

[5]. Donner, P, Enhanced self-citation detection by fuzzy author name matching and complementary error estimates, 2016, JOURNAL OF THE ASSOCIATION FOR INFORMATION SCIENCE AND TECHNOLOGY. 662-670.

[6]. 马海收等，基于ISI Web of Knowledge引证检索服务统计软件设计与实现. 情报杂志, 2012. 31(02): 第148-152+135页.

[7]. 马芳珍等，对CALIS查收查引系统的测试和应用效果评价. 大学图书馆学报, 2016. 34(02): 第97-102页.

[8]. 王学勤等，“查收查引报告自动生成系统”应用实践研究. 图书情报工作, 2014. 58(16): 第131-137页.

[9]. 刘彦君等，一种基于引文网络与科研合作网络的领域专家遴选方法[P]. 申请号:201911154798.4.

[10]. 唐向东等，一种申请人自我引用分析系统[P]. 申请号: 200910194764. 8.

[11]. Tsay, MY, The Relationship between Journal Self-citation and Other Scientometric Data for Some Subjects of the Social Sciences. Proceedings of the International Conference on Scientometrics and Informetrics, 2009.

[12]. James, H., To cite or not to cite: author self-citations and the impact factor. Scientometrics, 2012. 92(2).

[13]. John, P.A.I., A generalized view of self-citation: Direct, co-author, collaborative, and coercive induced self-citation. Journal of Psychosomatic Research, 2015. 78(1).

[14]. Dag, W.A., A macro study of self-citation. Scientometrics, 2003. 56(2).

[15]. Bart, T. and G.N. Wolfgang, The influence of author self-citations on bibliometric meso-indicators. The case of european universities. Scientometrics, 2006. 66(1).

[16]. Thijs, B ; Glanzel, W, The influence of author self-citations on bibliometric meso-indicators. The case of European universities. SCIENTOMETRICS, 2006(08).

[17]. 简书网. 精确率和召回率定义辨析 [EB/OL]. [2018-08-15]. https://www.jianshu.com/p/4434ea11c16c.

[18]. 百度百科网. f-measure [EB/OL]. [2021-10-28]. https://baike.baidu.com/item/f-measure/913107?fr=aladdin.

[19]. 徐东亮，高性能在线模式匹配算法研究, 2015, 哈尔滨工业大学. 第 121页.

[20]. 张丽霞与宋鸿陟，改进的近似模式匹配算法. 计算机工程与设计, 2011. 32(05): 第1820-1823页.

[21]. 秦建，孙秀锋与吴春明，“垃圾短信”监控的中文多模式模糊匹配算法. 西南大学学报(自然科学版), 2013. 35(03): 第168-172页.