中文生物医学文献主题标引中副主题词自动组配机制探讨①

李军莲　李丹亚　孙海霞　李　芳　冀玉静

（中国医学科学院医学信息研究所 北京 100020）

摘　要：简要介绍了当前国内外副主题词自动组配的研究现状，提出了适用于中文生物医学文献处理的基于拼图－统计相结合的副主题词自动组配实现方法，详细阐述分析了该混合策略方法的实现机制及实现效果，并指出后续提高和改进建议。

关键词：副主题词；副主题词组配；主题标引；自动组配

The Mechanism of MeSH Subheading Automatic Attachment for Chinese Biomedical Literature

Abstract: This paper briefly introduces the current research status of automatic MeSH subheading attachment both home and abroad. After systematic research, a "Jigsaw puzzle"-statistical method combined approach is proposed, which is suitable for dealing with Chinese biomedical literature. Moreover, the realization mechanism of this integrated method is analyzed in details and the corresponding results are evaluated. In addition, suggestions expected to improve the practical value of main heading/subheading pair recommendation are further raised.

Key words: subheadings; subheading attachment; MeSH indexing; automatic attachment

1. 引言

副主题词是用于对主题概念进行限定的一类词汇，强调主题概念的某些专指方面。主题标引中，通过副主题词与主题词组配，不仅可以提高揭示文献的专指性，而且能清晰反映主题概念间的关系，全面提升检索系统的准确率。

副主题词组配是生物医学文献主题标引中最常见的形式，约 90% 的文献在标引时应考虑组配合适的副主题词[1]。2002 年，中国医学科学院医学信息研究所研制的中文生物医学文献主题标引系统投入实际应用后，极大地提高了中国生物医学文献数据库（CBM）入库文献的标引效率和标引质量[2]。但该系统目前只能推荐游离主题词和副主题词，副主题词组配标引基本采用人工方式进行，很难满足中文生物医学文献快速增长的需要，是制约主题标引工作效率的瓶颈之一。

国外，美国国立医学图书馆（NLM）一直走在此项研究的前列，其自动标引项目于 2007 年前后将副主题词自动组配研究提上日程[3]，阶段研究成果已在其联机标引系统（DCMS）中得到了初步应用。通过向标引员提供副主题词组配推荐，在一定程度上提升了文献标引效率，促进了 Medline/PubMed 数据库的发展。

鉴于此，课题组在全面解析 NLM 医学文本标引工具（MTI）副主题词自动组配方法的实现机制和

① 本文得到中国医学科学院医学信息研究所基本科研业务专项"中文生物医学文献主题标引系统副主题词自动组配机制的探讨"（项目编号：R090212）和国家"十二五"科技支撑计划课题"科技文献信息知识对象自动标注与集成技术开发"（项目编号：2011BAH10B05-N）的资助。

实现效果[4]的基础上，结合中文生物医学文献主题标引的特点，提出了多种副主题自动组配实现方案，包括拼图法、统计法、UMLS 语义网方法及拼图－统计相结合的混合策略方法，并从实用可行角度对各方法进行了初步实现与评估，最终结果表明混合策略方法较好。本文基于拼图－统计相结合的混合策略对中文生物医学文献副主题词自动组配实现机制及其实现效果进行详细介绍。

2. 基于混合策略的中文生物医学文献副主题词自动组配实现机制

2.1 基本思路

混合策略综合吸纳了 MTI 拼图法和统计法的思想，如图 1 所示。基本思路为：首先根据各类映射表，推荐游离主题词、游离副主题词和既定主题词/副主题词，依据《医学主题词表》(MeSH)标引规则将游离副主题词按照可组配的主题词进行分组，然后基于构建好的专题语料进行统计学习，实现游离主题词与副主题的组配推荐。

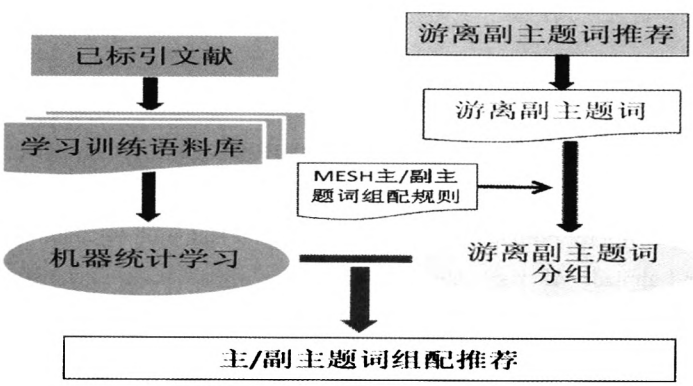

图 1 基于混合策略的副主题词组配推荐实现思路

2.2 实现步骤

图 2 是基于拼图－统计相结合的混合策略的副主题词自动组配系统实现流程图，主要包括 4 个步骤：语料统计训练、待标引文献主题初步提取、主/副主题词组配学习、结果筛选与输出。

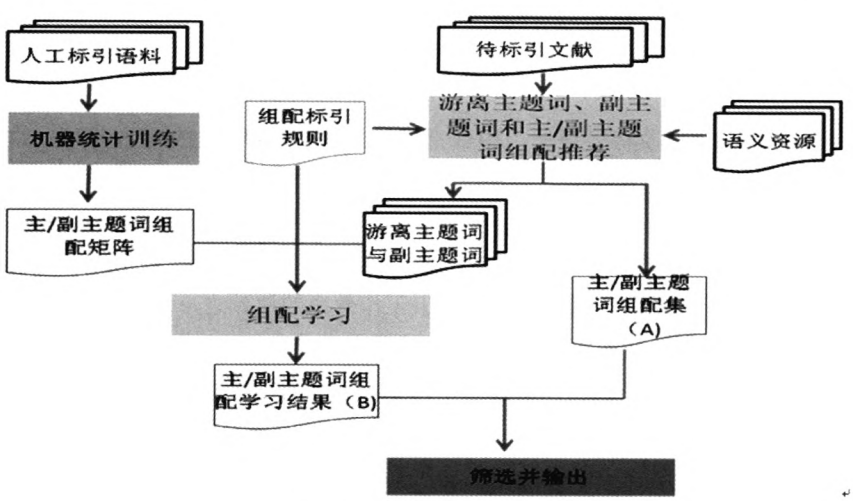

图 2 基于拼图法和统计的混合方法实现流程图

2.2.1 语料统计

基于构建好的专题人工标引语料进行主题词与副主题组配情况统计。统计结果存储于二维矩阵

中,行为 94 个副主题,列为语料中出现的主题词,值为副主题与主题词在语料中的组配频次。

2.2.2 待标引文献主题初步提取

根据长期积累的医学关键词语料、关键词 - 主题词映射表、关键词 - 副主题映射表、关键词 - 特征词映射表和 MTI 标引规则库等语义资源,通过分词、加权等技术,实现文献游离主题词、游离副主题和既定主/副主题词组配抽取。

2.2.3 游离副主题词与主题词组配学习

首先根据 MTI 标引规则,以主题词为单位,对游离副主题词进行分组。因一个副主题可以与多个主题词组配,所以一个副主题词可以属于多个组。以主题词 H 为例,其对应的副主题组 SH 中为当前文献中所有可与 H 组配的游离副主题词。

然后对每个主题词可组配的副主题词进行学习。同样以主题词 H 为例,学习过程为:首先依次判定 SH 中的游离副主题词是否出现在主/副主题词二维矩阵中主题词 H 对应的副主题词集合中;如出现,再判定是否满足其它权值条件,满足则存于主/副主题词组配集合 B 中。如 SH 中的所有游离副主题词都未出现,取矩阵中最高频次的副主题词与 H 组配,存于 B 中。

2.2.4 结果筛选与输出

对既定主/副主题词组配抽取结果与 B 中的学习结果进行融和,按既定过滤规则进行筛选与输出。

3. 中文生物医学文献副主题词自动组配实现效果分析

为进一步验证该思路、方法的科学性与可行性,课题组开展了如下实验:按照设计思路初步实现了中文生物医学文献副主题词的自动组配,并对随机抽取的部分实现结果进行分析,从副主题词推准率(Precision)和推全率(Recall)两方面对实现效果进行评价。

3.1 实验方案

本实验主要分为两个阶段:一是基于原有的中文生物医学文献主题标引系统实现“中文生物医学文献副主题词的自动组配”,二是对实现结果进行分析评价。为方便后期分析评价的进行,拟从“标引训练集”中抽取部分数据进行“副主题词自动组配”实现,将处理结果与相应的人工标引结果进行比较分析。

3.2 实验数据集

在本实验中,涉及两种实验数据,一为标引训练集,一为实验结果分析数据集。

3.2.1 标引训练集

标引训练集用于统计学习主题词与副主题词的组配规律,其标引质量高低、学科覆盖全面与否直接影响统计学习训练的效果。根据 CBM 数据库的标引要求,最终遴选出标引深度较深、标引质量较好、涵盖学科较全的 45160 篇文献作为标引训练集,具体如下:

(1)期刊分布:以中华系列、中国系列期刊为主,同时选取了部分大学学报及高质量交叉学科期刊。

(2)领域分布:以医学综合、临床综合为主,尽量涵盖医学各研究领域,具体分布详见表 1。

(3)年代分布:2003 年 3141 篇; 2006 年 15789 篇; 2008 年 25742 篇。

3.2.2 实验分析文献集

选取标引训练集中的前 2 万篇文献进行“副主题词自动组配”效果测试,从中随机抽取 150 篇文献作为实验分析数据集进行副主题词自动组配效果分析。

3.3 评估方法

通常,对于自动标注效果的评价主要有两个指标:推全率和推准率。在本实验中,拟借用这两个指标,将副主题词自动组配结果与人工标引结果及副主题词游离推荐结果进行对比分析,从而评估其科学性和可行性。

表 1　标引训练集学科领域分布

学科领域	文献量(篇)	文献百分比(%)
医学综合	12198	27
环境医学、卫生管理	2956	6.5
中医、中药	1701	3.8
基础医学	2887	6.4
临床综合(含内、妇、儿、肿瘤、口腔、眼科、特种医学等)	20974	46.4
外科学	6230	13.8
药学	1271	2.8
兽医学	164	0.4
交叉学科	1789	4

具体地讲,推全率用于评价自动推荐出来的主副组配结果与人工标引的吻合程度,而推准率则用于反映自动推荐出来的主副组配的准确程度,二者的计算公式如下:

推全率 = 自动标引与人工标引相同的主副组配数量/人工标引主副组配数量;

推准率 = 自动标引与人工标引相同的主副组配数量/自动标引主副组配数量。

3.4　实验结果分析

考虑到与现有中文生物医学文献主题标引系统的兼容性与可整合性,本实验通过 C + + 编程来实现主题词和副主题词的自动组配。对测试集 2 万篇文献完成"副主题词自动组配"处理后(表 2 是部分测试结果),随机从中抽取 150 篇文献作为实验分析数据,进行副主题词自动组配效果的判定分析,具体如下:

表 2　测试结果示例

文献 UI	2004126830
文献标题	豚鼠前庭椭圆囊器官在庆大霉素损伤后离体培养的细胞增殖
游离推荐副主题词	/损伤; /细胞学; /方法; /解剖学和组织学; /教育
游离推荐主题词	*庆大霉素类; *球囊和椭圆囊; *创伤和损伤; *溴脱氧尿苷; 前庭, 迷路; 毛细胞, 听觉; 细胞, 培养的; 细胞增殖
自动主副组配推荐	MYM 创伤和损伤/*病理学; MYM 庆大霉素类/*毒性; @球囊和椭圆囊/*细胞学; *溴脱氧尿苷; @毛细胞, 听觉/细胞学; @前庭, 迷路/细胞学; MYM 细胞, 培养的/药物作用; MYM 细胞增殖/药物作用
人工标引	毛细胞, 听觉/*药物作用; 庆大霉素类/*毒性; 球囊和椭圆囊/*细胞学; 细胞, 培养的; 细胞增殖/药物作用; 溴脱氧尿苷
文献 UI	2008431157
文献标题	慢性阻塞性肺疾病的肺减容术治疗
游离推荐副主题词	/治疗; /外科学; /方法; /并发症; /死亡率; /药物疗法
游离推荐主题词	*肺疾病, 慢性阻塞性; 肺切除术
自动主副组配推荐	@肺疾病, 慢性阻塞性/*外科学; @肺切除术/方法
人工标引	肺疾病, 慢性阻塞性/*外科学; *肺切除术
文献 UI	2008431016
文献标题	活性氧在慢性缺氧性肺动脉高压发病机制中的作用
游离推荐副主题词	/生物合成; /解剖学和组织学; /化学
游离推荐主题词	*缺氧; 慢性病; *活性氧; *高血压, 肺性; *肺动脉; NADP; 肺; 动脉
自动主副组配推荐	MYM 肺动脉/*外科学; MYM 高血压, 肺性/*病理生理学; MYM 活性氧/*代谢; MYM 缺氧/*病理学; NADP; @动脉/解剖学和组织学; @肺/解剖学和组织学; 慢性病
人工标引	活性氧/*代谢; 慢性病; 缺氧/并发症; 高血压, 肺性/*病理生理学; 高血压, 肺性/病因学

(1)本研究所提出的基于混合策略的中文生物医学文献副主题词自动组配方法不仅能覆盖现有的 94 个副主题词,而且具有较好的推全率和推准率(推全率为 61.5%,推准率为 48.3%,见表 3)。

(2)与原有自动标引系统游离副主题词推荐效果(推全率为 43%,推准率为 24.7%,见表 4)相比,推全率和推准率均有比较明显的提升。

(3)组配推荐效果基本达到了美国国立医学图书馆 MTI 文本标引工具副主题词自动组配的实现水平[5-6]。与西文的各种处理方法相比,推准率虽低于"后处理规则法",但却高于其他几种方法,包括综合法(见表 5)。

表 3　副主题词自动组配推荐效果分析

文献 ui	人工标引主副组配总量(个)	自动标引主副组配		推全率	推准率
		总量(个)	与人标相同量(个)		
2004126830	4	6	3	0.75	0.5
2004126829	4	5	0	0	0
2004126828	6	8	3	0.5	0.375
2004126470	6	6	2	0.333333	0.333333
2004126438	4	6	3	0.75	0.5
2004125940	1	1	1	1	1
…	…	…	…	…	…
平　　均				0.615	0.483

表 4　原有自动标引系统游离副主题词推荐效果分析

文献 ui	人工标引主副组配总量(个)	游离推荐副主题词		全率	推准率
		总量(个)	与人标相同量(个)		
2004126830	4	5	1	0.25	0.2
2004126829	4	4	1	0.25	0.25
2004126828	6	5	2	0.333333	0.4
2004126470	6	9	3	0.5	0.333333
2004126438	4	5	3	0.75	0.6
2004125940	1	4	1	1	0.25
…	…	…	…	…	…
平　　均				0.43	0.247

表 5　中、西文文献副主题词自动组配实现效果对比表

文种	方法名	副主题词量	实现效果(%)	
			推准率	推全率
中文	拼图 - 统计混合法	94	48.3	61.5
西文	基于词典的方法(DIC)	83	26	31
	MTI 法	82	24	13
	后处理规则法(PP)	19	58	5
	自然语言处理规则法(NLP)	20	38	2
	PubMed 相关文献推荐法(PRC)	83	35	54
	综合法(至少两种方法 + 过滤)	83	44	29

4. 结论与展望

副主题词组配标引不仅可以提高揭示文献的专指性，而且能清晰反映主题概念间的关系，全面提升检索系统的准确率。实验初步表明，本研究提出的基于混合策略的中文生物医学文献副主题词自动组配方法具备一定的科学性、可行性。该方法不仅能全面覆盖现有的副主题词，而且具有较好的推全率和推准率，通过进一步优化完善可逐步考虑投入实际应用。

同时，通过分析自动主副组配误推、漏推的原因，课题组也注意到混合策略副主题词自动组配方法在设计和实现中有些环节考虑得还不全面，科学性、严谨性有待进一步加强，今后还需从如下方面予以提高：

（1）开展医学文献标引专题语料构建模型研究，最大限度提高组配学习效果。

统计训练、学习人工标引结果是基于混合策略的副主题词自动组配方法的主要思想之一，故标引训练集构建的科学与否至关重要，直接影响到主副组配训练学习的效果。在本研究中，虽也从标引深度、标引质量、涵盖学科三方面考虑构建标引训练集，但并没有从量上进行有效控制，譬如在目前的训练集中，近 3/4 为医学综合、临床综合方面的文献，中医中药、药学、兽医学等方面的文献均不足 5%（见表 1），这也正是实验结果中中医药方面副主题词自动组配效果不好的主要原因。在后续的研究中建议分领域构建标引训练集，有效控制语料规模和质量，最大限度提高主副组配训练学习效果。

（2）进一步解析系统现有标引机制，全面提高各环节实现效果。

本研究最初设计是基于现有主题标引系统副主题词游离推荐功能来探讨副主题词的自动组配机制，故主副自动组配实现过程中不可避免要受到游离主题词、游离副主题词推荐准确全面与否的影响。尽管在实验之前已经对“副主题词游离推荐”的思路进行了优化，但由于缺少进一步验证，在实验中仅用到扩充优化后的“关键词—副主题词映射表”，基于“主题词—副主题词”规则库的副主题词推荐方法暂时没有考虑。由此可见，要最终提升主/副自动组配的效果，必须全面考虑主题标引的各个环节，仅仅停留在“组配”机制的探讨上是远远不够的。

（3）继续考虑发挥“规则”在主副自动组配中的作用。

鉴于高级、完备规则建立的复杂性，本研究最终提出的基于混合策略的副主题词自动组配方法将重点放于统计学习法和拼图法上，暂时弱化了基于“UMLS 语义关系”的规则法的应用。虽也得到了较好的实验效果，但错误推荐、无效推荐仍占相当比例，这无疑为标引人员带来了另一方面的负担。故在后续研究中，还应考虑发挥“规则”在主/副自动组配中的作用，通过简单、有效的规则控制，切实提高主副组配的准确性，减少错误推荐和无效推荐。

参考文献

[1] 肖晓旦，张士靖. 医学文献主题标引[M]. 北京：高等教育出版社，2008：60.

[2] 钱庆，胡铁军，李丹亚，等. 中文生物医学文献主题标引系统的研究[J]. 医学情报工作，2002，(2)：84-86.

[3] Alan R. Aronson, James G. Mork, Feancois-Michel Lang, Willie J. Rogers, Aurelie Neveol. NLM Medical Text Indexer: A Tool for Automatic and Assisted Indexing. http://ii.nlm.nih.gov/resources/ii-bosc08.pdf (2010-03-27).

[4] 孙海霞，李军莲，李丹亚，等. MTI 副主题词自动组配标引机制解析[J]. 医学信息学杂，2011，32(5)：74-77，90.

[5] Névéol A, Shooshan SE, Humphrey SM, Rindflesch TC and Aronson AR. Multiple approaches to fine-grained indexing of the biomedical literature[J]. Proc. PSB, 2007(12): 292-303.

[6] Aurélie Névéol, Sonya E. Shooshan, James G. Mork, Alan R. Aronson. Fine-Grained Indexing of the Biomedical Literature: MeSH Subheading Attachment for a MEDLINE Indexing Tool[J]. AMIA 2007 Symposium Proceedings: 553-557.

中文生物医学文献主题标引中副主题词自动组配机制探讨

作者： 李军莲， 李丹亚， 孙海霞， 李芳， 冀玉静
作者单位： 中国医学科学院医学信息研究所 北京 100020