非专利文献同义词库的维护方法研究

专利检索咨询中心 鄢燕 章洪流 李小生

摘要: 非专利文献同义词库的构建具有重要意义。本文分析总结了当前非专利文献数据深加工同义词库的现状和问题,说明了同义词库定期维护的必要性,对同义词库进行了整理。整理后的非专利文献同义词库,克服了单纯靠计算机处理的缺陷,提高了准确度,能够更好的为数据深加工和检索提供优质服务。

关键词: 非专利文献 同义词库 整理 维护 数据加工

引言

同义词标引是非专利文献数据加工的一项重要内容，其意义在于加工形成的同义词库一方面可以为非专利文献数据加工本身提供支持，避免化学结构重复加工，数据重复录入，提高方剂信息、IPC 加工效率；另一方面可以为非专利文献的检索提供便利，通过同义词库扩展关键词，能提高检索的查全率和查准率 [1]。但是，非专利文献撰写灵活,存在一词多义、略缩语、俗语、简化字等情况。随着

数据的加工量不断增加，同义词库中词条数量增长迅速，同义词库的利用不可避免的出现了一些问题。如何保证同义词库发挥稳定有效的作用，成为数据加工一个新的研究方向。

一、非专利文献数据深加工同义词库的现状

目前，我们已经对800多种医药类非专利期刊文献进行了数据深加工。随着文献量不断增加，同义词库更新迅速。仅2010年9月至2011年10月期间，加工收录的同义词就有三万七千多条。与中国药物专利数据库相比，非专利文献同义词库收录的同义词条内容要丰富得多。以“中药菝葜”为例，在中国药物专利数据库中搜索“菝葜”，共13条结果，而搜索非专利文献同义词库，有24条结果，见表1。

表1 专利数据库与非专利文献数据深加工同义词库对比

中国药物专利数据库	非专利文献数据深加工同义词库
菝葜、小叶菝葜、短梗菝葜、防已叶菝葜、粘鱼须菝葜、武当菝葜、鞘菝葜、蜜疣菝葜、草菝葜、抱茎菝葜、短柄菝葜、粉背菝葜、白菝葜	菝葜、小叶菝葜、短梗菝葜、防已叶菝葜、柔毛菝葜、鞘柄菝葜、密刺菝葜、尖叶菝葜、华肖菝葜、华东菝葜、暗色菝葜、短柱肖菝葜、光叶菝葜、菝葜属、长托菝葜、小果菝葜、西南菝葜、穿鞘菝葜、粗糙菝葜、黑叶菝葜、托柄菝葜、梵净山菝葜、肖菝葜、黑果菝葜

丰富的词库数据建立在内容浩瀚的非专利文献基础上。非专利文献同义词的标引遵循尊重原文的原则，而单篇文献提取的同义词组往往不全面，需要通过计算机自动处理对相同同义词元素进行链接，才能形成较为完整的同义词条。但是，计算机自动处理存在一定的局限，不能识别由于文献作者撰写习惯、一词多义、略缩语、标引错误等原因导致的数据差别，从而产生词组链接错误或不能链接的问题。如表2，词组1-3由于存在相同的元素“SDS”故被计算机链接成一个同义词条，而实际上这3组词组分属药学、医学、化学不同领域，意义完全不一样，不应作为同义词被链接；词组4-6对应的化合物均为“5-氟尿嘧啶”，但由于3个词组间没有完全相同的字符，计算机无法将这3个词条进行链接。上述问题的积累，

导致同义词库日渐繁冗，如果不进行适当的处理，同义词库将无法持续地发挥稳定精确的作用。

表 2　链接错误和不能链接示例

词组 1	十二烷基磺酸钠；SDS
词组 2	扩展性抑制波；spreading depressions；SDs
词组 3	自主神经受累；SDS
词组 4	5- 氟尿嘧啶；5-Fu
词组 5	氟尿嘧啶；fluorouracil
词组 6	5 氟尿嘧啶；5FU

二、同义词库的人工干预维护

基于以上原因，同义词库不可避免地需要人工干预，很有必要定期组织专门人员对同义词库进行库内的纠错和整理规范，以便加工人员在加工过程中，对存在疑问或不确定的同义词随时检索查询，以提高加工准确率和检索查全查准率。

同义词库的维护主要分为 4 种处理方式:即删除、合并、修改和保留。对不需要标引的同义词予以删除；对不同表达形式、重复的同义词进行合并；对标引或链接错误及标引不规范的同义词进行必要的修改；对核实正确的同义词，直接保留入库。

（1）删除

删除的同义词条大致分为以下几种情况。①一些解释性或说明性描述的词句被标引员加工为同义词条,如:原发性肝癌(癌症之王);无形之痰(痰浊)；百会（三阳五会）；众生丸（中药抗生素）……②标引早期，一些机构及非医药类的词组被标引，如：中国典型培养物保藏中心（CCTCC）；质控（QC）；校正均方根误差（RMSEC）；相对标准差（RSD）……③没有意义，简单中英文对照被标引，如一般状况记分标准（KPS）；叶（leaf）；样品（sample）；对照组（control group）……④含单个字或字母的同义词,如利福喷丁(T);肾上腺素(E)。

（2）合并

将不同表达形式、重复的同义词进行合并,具体整理情况如表 3 示例，合并处理主要集中在中药和化合物方面。中药品种繁多，名称复杂，同名异物、同物异名现象常见，同音字、简化字使用泛滥，还存在古名、别名、地方书名等现象，而这些情况在非专利文献中尤为明显。化合物方面，同一化合物的中英文对照和缩写有多种形式，且每个作者的书写表达习惯也

不同，这使得同一化合物词条有可能被多次标引。

表 3 合并情况

	整理前	整理后
合并	薄荷；Mentha haplocalyx Briq.	薄荷；薄荷；卜荷；Mentha haplocalyx Briq.
	薄荷；卜荷	
	生薏苡仁；生苡仁	生薏苡仁;生苡仁;薏苡仁;薏米;苡米；苡仁；生苡米仁；生苡米
	薏苡仁；薏米	
	苡米；苡仁；生苡米仁；生苡米	
	5-氟尿嘧啶；5-Fu	5-氟尿嘧啶；5-Fu；氟尿嘧啶；fluorouracil；5 氟尿嘧啶；5FU
	氟尿嘧啶；fluorouracil	
	5 氟尿嘧啶；5FU	

（3）修改及保留入库

对标引或链接错误的同义词条以及标引不规范的同义词条进行修改，将修改后的同义词条重新入库。核实正确的同义词直接保留入库。示例如表 4。

表 4 修改及保留入库情况

	整理前	整理后
修改	盐酸黄连素片；Berberine Hydrochloride Tablets；盐酸小檗碱片；黄连素；小檗碱;盐酸小檗碱;盐酸黄连素;Berberine	盐酸黄连素片;盐酸小檗碱片;Berberine Hydrochloride Tablets；盐酸小檗碱；盐酸黄连素
		黄连素；小檗碱；Berberine
	流行性出血热休克；EHF	流行性出血热；EHF
保留入库	叶牡丹；Leontice robustum	叶牡丹；Leontice robustum

三、整理后的同义词库

在尝试对 37302 条同义词进行整理后，词条精简为 18904 条。从空间意义来说，节省了存储资源；从利用效果上看，不仅对数据加工人员有了更准确的参考价值，能提高加工准确

率，也能提高同义词检索的查全率和查准率。举例说明，以“生薏苡仁”为关键词，用我部门开发试用的“非专利文献检索系统”进行简单检索[2]，共有 11 条检索结果，如图 1 所示。利用整理前的同义词库进行“同义扩词”检索，检索的关键词包括“生薏苡仁 + 米仁 + 生米仁 + 生苡仁 + 生薏米 + 生薏米仁 + 生薏仁 + 生薏苡仁米 + 苡仁苡米 + 薏米 + 薏仁 + 薏仁米”，能检到 41 篇文献，如图 2 所示。利用整理后的同义词库进行“同义扩词”检索，检索的关键词为“生薏苡仁 + 薏苡 + 生苡米仁 + 生苡米 + 薏苡仁 + 米仁 + 生米仁 + 生苡仁 + 生薏米 + 生薏米仁 + 生薏仁 + 生薏苡仁米 + 苡仁 + 苡米 + 薏米 + 薏仁 + 薏仁米”，检索结果扩展为 195 篇，如图 3 所示。

通过整理后的同义词库扩展检索，命中的文献量比原同义词库扩词检索要多得多，且经过浏览分析，其检索结果的准确度高，命中文献具有相当的参考价值。可见，同义词库以及同义词扩展功能的使用对非专利文献检索“查全查准”起到了不可或缺的重要作用，而定期整理维护正是保证同义词库发挥持续稳定作用的必要手段。

图 1 简单检索

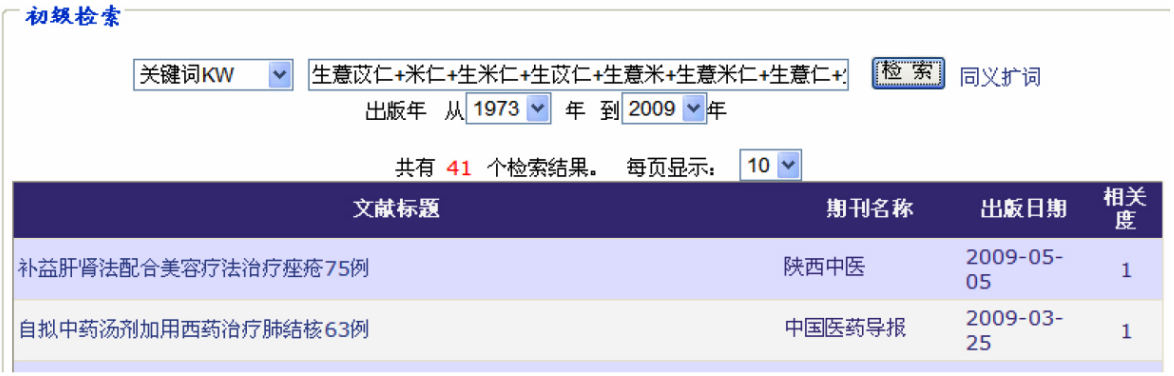

图 2 原同义词库“同义扩词”检索

图 3 整理后的同义词库“同义扩词”检索

四、结语

从以上数据可以看出，非专利文献同义词库的建立，在现有“中国药物专利数据库”基础上，极大地丰富了词汇量，对数据加工本身而言具有很好的参考借鉴价值。经过加工标引的“非专利文献检索系统”是非专利文献检索的有效手段，而定期组织专门人员对同义词库进行库内的纠错和整理规范，对减少检索噪音，提高文献查全查准率是很有必要的，是检索系统在不断更新中的必需维护手段。相信非专利同义词库会在未来的专利审查工作中发挥更加重要的作用。

（专利检索咨询中心 杨晓春 李娜 审校）

参考文献

1. 颜平辉、孙亮、章洪流．“非专利文献同义词库构建与应用研究”．《数据加工通讯》2012 年第 6 期第 16-27 页。
2. 张旻、孙亮、庄莹、李林福．“非专利文献数据加工同义词标引规范化及其在检索中的应用”．《数据加工通讯》2009 年第 19 期第 10-16 页。