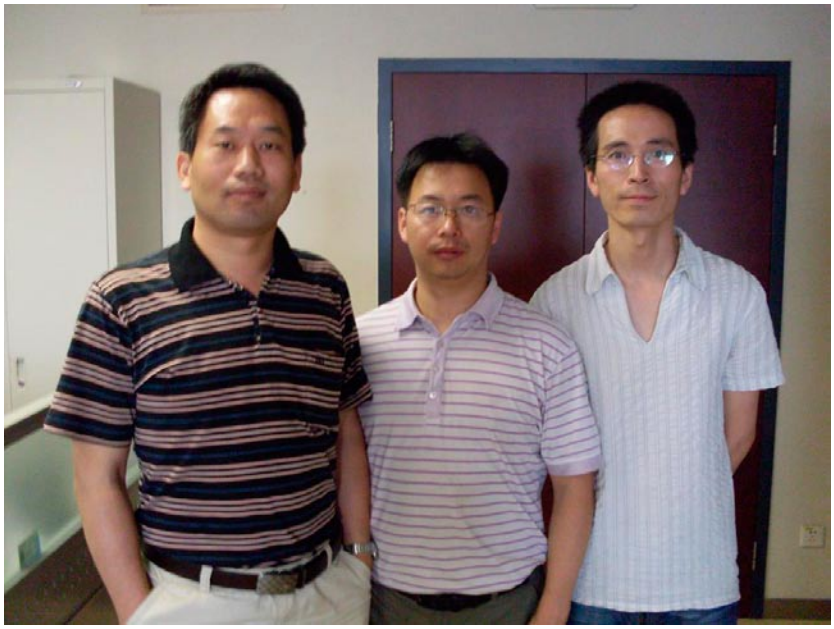
非专利文献同义词库构建与应用研究

专利检索咨询中心 颜平辉 孙亮 章洪流



摘要: 本文分析了非专利文献数据加工中同义词库构建的重要意义、现状和同义词库存在的问题,提出了一套人机协作构建同义词库机制。通过精心设计的数据存储结构和同义词连接算法,使计算机具备了学习能力,在数据加工过程中智能化水平不断上升,自动构建同义词库准确度越来越高,人工干预工作量日趋变小,以较小代价获得了准确可用的同义词库,并将同义词库应用于数据加工和文献检索中,结果证明其作用是显著的。

关键词: 非专利文献 同义词库构建 应用研究

一、研究背景

同义词加工是非专利文献数据加工中的一项重要内容，加工形成的同义词库为非专利文献数据加工本身提供支持，在避免化学结构重复加工，减少数据重复录入，提高方剂、IPC加工效率方面均有重要作用。更为重要的是，在非专利文献检索时扩展关

键词，从而提高查全率和查准率，为专利审查提供可靠支持。

目前，在非专利数据加工中，加工工作是以单篇文献为基本加工对象，提取对专利审查有益的信息，如同义词、方剂、化学结构等信息进行标引。就同义词加工而言，按照尊重原文的加工原则，只对原文出现的同义词组进行标引，这样，单篇文献提取的同义词组通常是不全面的，在较大数据范围考虑所加工的同义词时，就会出现实为一组的同义词被分割成多个不完整、有重复且不相关的词组。直接利用加工形成的同义词库进行扩展查询时不仅查询结果不全面，而且会出现查询结果不一致的现象。因此，对加工的同义词组进行整理、连接、重构和维护，形成一个相对完整和正确的同义词库是一项非常重要而有意义的工作。

计算机自动处理与人工干涉相结合的方法构建同义词库，计算机将相同同义词元素连接形成较为完整的同义词。人工干涉完成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连接错误或不能连接的词组。本文研究重点在于构建一套人机协作机制，共同构建、维护同义词库，使计算机在构建同义词库的过程中持续学习，智能化水平不断提高，同义词连接准确度越来越高，人工干预工作量越来越小；并研究如何将同义词库应用于数据加工和文献检索中。

二、同义词库构建与维护

同义词库构建与维护主要依赖于数据存储结构，原始同义词表和同义词基表。原始同义词表存储了标引过程中针对单篇文献提取的同义词组信息；同义词基表存储了经过整合的同义词信息。在标引过程中，每标引一组同义词，系统将其存储于原始同义词表中，同时去检查该同义词在同义词基表的情况，自动进行连接并作相应处理，动态构建同义词基表。标引人员对自动构建的同义词库进行定期维护，进行正确性确认，系统将学习维护过程中标引人员用到的专业知识，从而为此后连接同义词提供支持，实现更高的连接准确率，降低确认工作量。

（一）数据存储结构

原始同义词表，主要存储单篇文章的同义词组，并记录该同义词组是否经过整合的状态信息。可通过DOI查询该文章的所有同义词组，也可用某个词查出包含该词的所有同义词组和DOI。其表结构和样例数据如表1所示，各字段说明如下：

表 1 原始同义词表

DOI	SynOrder	PartSynContent	Domain	CheckStatus
…	216961	羟丙基－β－环糊精；HP-β-CD	医药	yes
…	502036	硫唑嘌呤；Azathioprine；AZa	医药	no
…	804392	甲氧苄胺嘧啶；Trimethoprim；TMP	医药	yes
…	866885	甲氧苄啶；甲氧苄胺嘧啶；TMP	医药	yes

DOI(Digital Object Identifier):数字对象标识符，唯一标识一篇非专利文献。通过DOI可联接查询原文信息、关键词、方剂、化学结构、IPC和标引人员、质检情况等过程的所有信息。

SynOrder：一条同义词组在该文章所有同义词组中的序号，该序号与标引单中序号一致，可通过DOI和SynOrder定位任意一篇文章的某一条同义词组。

PartSynContent：一条同义组的具体内容，同义词间用分号分隔。

Domain：该同义词组所在的学科领域，如化学、医药。

CheckStatus:取值为‘no’和‘yes’，指示该同义词组是否已经整合到同义词基表。

同义词基表，主要存储了不同领域经过计算机整合的同义词组。这些同义词组包括已经确认和未确认的，确认结果分为正确，无意义和错误三类。同义词基表不仅存储了未确认的同义词和确认为正确的同义词，还存储了经确认为错误和无意义的同义词，其目的在于将专业人员的判断结论存储下来。无论同义词组正确与否，都是有用的结论，并将其转化为计算机可理解的知识，再次遇到类似情况如无意义的同义词时，可提醒标引员不需要重复标引，已经发现连接错误的，将寻找新的连接方法。此外，还可有效地防止同义词基表重复确认。对同义词基表的应用则根据不同的应用场景提取不同状态的同义词，同义词正确性检查时则查找确认为无意义的同义词，找到说明有标引有误；化学结构标引时则查找正确和未确认的同义词，保证最大限度的查找到已绘结构，在线非专利文献检索系统则使用正确的同义词，实现一表多用，也避免了计算机和人的重复工作。其表结构和样例数据如表2所示,各字段说明如下:

SynID:唯一标识一条同义词组；

FullSynContent：经整合后的一条同义词组具体内容，同义词间用分号分隔，为便于查询匹配，同义词组最前也加上分号。

Domain：该同义词组所在的学科领域，如化学、医药。

UpdateStatus：更新状态，取值

为：‘new’，‘yes’和‘no’。‘new’表示新增需要确认，若是计算机自动插入，则只在一篇文献中出现过；‘yes’表示更新过需要确认，是通过多个同义词组连接而来，通常出现在多篇文献；‘no’表示已经经过人工确认，同义词更新后，其状态变为‘yes’，指示该同义词组需要人工再次确认其正确性。

ConfirmResult：经人工确认后的结果，取值为：正确、无意义、错误，默认为未确认。

此外还包括 UpdatePerson，UpdateTime，ConfirmPerson，ConfirmTime 字段记录同义词组确认人员和时间信息，以便工作任务分配、统计、追踪。

表 2 同义词基表

SynID	FullSynContent	Domain	Update Status	Confirm Result
16176	；硝洛地平；2,6- 二甲基 -4-(2- 氯 -3- 硝基苯基)-1,4- 二氢吡啶 -3,5- 二羧酸甲酯；	医药	no	正确
12004	；中国典型培养物保藏中心；CCTCC；	医药	no	无意义
13007	；丹参；大红袍；红根；地黄；Salvia miltorrhizaBge; 金状元；白状元；血参根；	医药	no	错误
16938	；盐酸黄连素片；Berberine Hydrochloride Tablets；盐酸小檗碱片；黄连素片；	医药	yes	未确认
18839	；左酮洛芬；levoketoprofen；	医药	new	未确认

（二）自动构建方法

同义词库构建与标引同时进行，当标引单保存成功时，此时该标引单所有同义词已存到原始同义词表，字段 CheckStatus 值为‘no’，指示需要整合到同义词基表。系统会启动一个线程，负责同义词整合所有工作，该线程在后台执行，标引员意识不到同义词整合工作的存在，主要工作包括：

从原始同义词表提取该标引单的所有同义词组；

逐条检查同义词基表包含该同义词组情况（全部包含、部分包含、不存在）以及人工确认的结果信息，以此判断该同义词组是否需要更新同义词基库；

如果需要更新，则更新同义词基表，并修改同义词基表相应条目状态；

根据更新情况，修改原始同义词表 CheckStatus 字段值。

此过程中关键在于如何将单条同义词组加入到同义词基表中，该过程如图 1 所示，

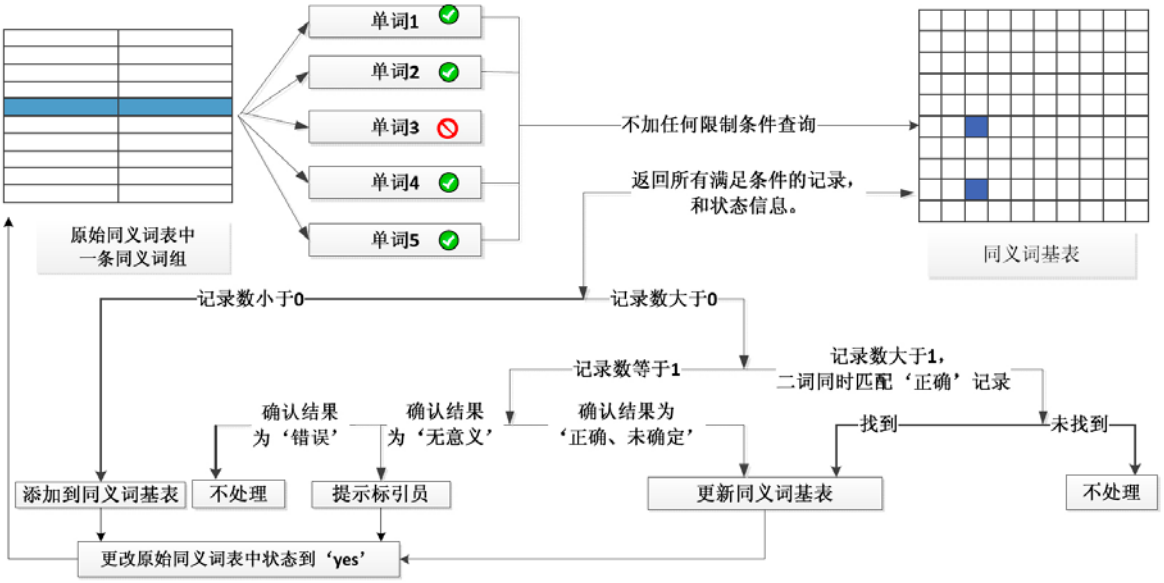


图 1 同义词基表构建过程

第一步，进行同义词拆分与过滤。将分号标识的同义词字符串转换为数组，过滤掉英文字符长度小于 4 的同义词元素。主要原因在于不同词语缩写后 4 字符以下相同的可能性极高，导致计算机将不相关的词组连接起来。过滤字符长度主要根据实际经验获得，可根据具体情况不断调整。

第二步，用前一步的结果，不加任何限制条件对同义词基表进行查询，取回所有满足条件的结果。查询结果应包括同义词 ID，内容，更新状态和确认结果信息。

第三步，查询结果分析和处理。

查询结果记录数等于 0，说明同义词基表中不存在该同义词组，可直接添加到同义词基表中。同义词基表中将添加的条目更新状态置为‘new’，表明新添加，同时将确认结果置为‘未确认’，表示需要人工确认，还需更新原始同义词表中 CheckStatus 值为‘yes’，表明已对该词进行过整合。

查询结果记录数等于 1，分析确认结果信息，如果为‘错误’，说明用一个同义词元素进行连接的方式被确认为错误的，除此之外同义词基表中没有任何可以进行连接的同义词组，所以不做任何处理，原始同义词表中 CheckStatus 值不变，以示需人工处理；确认结果信息为‘无意义’，说明该同义词不需要标引，也不需要再次进入同义词基表，所以对同义词基表不作处理，仅更新原始同义词表中 CheckStatus 值为‘yes’，表示已对该词进行过整合；确认结果信息为‘正确’或‘未确定’，说明在同义词基表中找到了一条可以连接的同义词组，处理方式为先检查同义词基表中

的同义词组是否完全包含待插入的同义词，如果是，则仅更新原始同义词表中 CheckStatus 值为 ‘yes’，如果不是，则将未包含的部分添加到同义词基表中，同时将同义词基表更新状态置为 ‘yes’，并将确认结果置为 ‘未确定’，表明该同义词组已被系统自动更新，需人工进行确认，再将原始同义词表中 CheckStatus 值为 ‘yes ’。

查询结果记录数大于 1，说明同义词基表中有多条可以进行连接的同义词组。这些同义词组通常是人工拆分的结果，既包括拆分前错误的同义词组，也包括拆分后形成的多个正确的同义词组。此时，如果按照单个同义词元素进行连接，由于无法确定该连哪一条同义词基表的记录，显然是不行的。采用的方法是，如果查询记录的确认信息为 ‘正确’ 或 ‘未确定’，用两个同义词元素同时对查询结果信息匹配，从而找到应该连接的那一条，再更新同义词基表和改写原始同义词表的状态信息，如果找不到，则不做任何操作，保持原始同义词表中 CheckStatus 值不变，以示需要人工处理。

（三）同义词库维护

同义词库维护工作主要包括两个方面，一是对所有新添加和更新过的同义词基表数据进行确认，判断其正确性，二是对原始同义词表中系统无法自动连接的同义词组进行人工干预生成同义词组。

同义词组正确性确认和修改，定期遍历同义词基表中更新状态为 ‘new’ 和 ‘yes’ 的同义词组，逐条判断，其结果分为三种类型：正确，无意义，错误。将标记为正确的同义词用于检索系统；标记为无意义的同义词主要用于提醒标引员，根据标引规则，不需要标引此类同义词；标记为错误的同义词主要用于系统自动构建同义词库时，告知系统用此种方法连接的同义词是错的，改用其它方法连接。通过人工确认结果信息，计算机也具备了专业人员的部分智能，在不断的维护过程中，计算机判断同义词的智能水平越来越高。

为了方便确认同义词组，需要提供该同义词组来源详细信息，其方法如图 2 所示，将同义词组字符串拆分同义词元素数组，单个同义词元素在原始同义词表中查找获得该同义词元素的详细标引信息，还可通过 DOI 获得原文、标引单、标引员信息。

对于错误的词条，需要分析错误原因，进一步处理。错误主要来源为标引错误和计算机自动连接错误。对于前者，可更正原始同义词表中标引错误或将其标记为不参与同义词连接的词条，更正同义词基表中错误词条后重新生成基表同义词组；对于计算

机连接错误，保留同义词基表中错误词条，以提示计算机在遇到该类同义词是不要用此方法连接，同时将错误词条拆分为多个正确的词条。通过确认和修改，同义词基表中的错误词条只剩下计算机连接错误的词条，提示计算机以后不要用此方法连接了，换用其它方法。

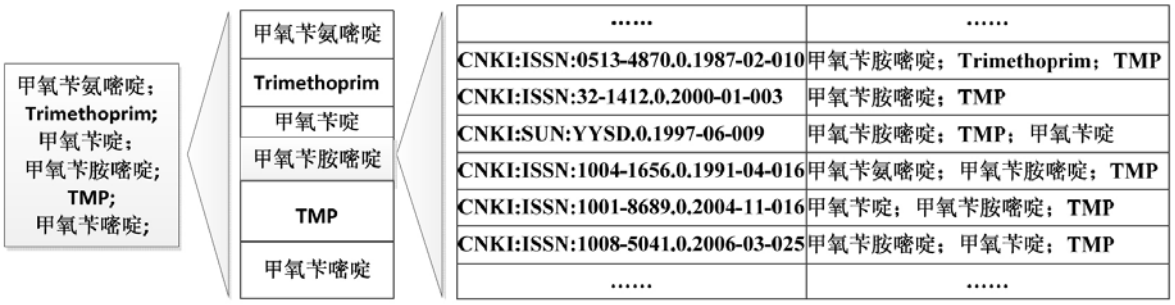


图 2 同义词来源信息查找方法

人工干预生成同义词组。是对原始同义词表中计算机未能处理的词条进行补充性处理，以保证所有标引的同义词都整合到同义词基表中，其关键是如何从原始同义词表中找出与每条同义词组可能连接的所有同义词，以便人工判断。其算法如下：

设置栈变量 SynStack 和字符串变量 FullSyn；

将同义词组分割为同义词元素，逐元素与 FullSyn 比较，如果 FullSyn 中不包含该元素，将该元素添加到 FullSyn, 同时入栈；

出栈，取出一个同义词元素，在原始同义词表中查询包括该同义词元素的所有同义词组，执行第 2 步；

如果栈空，执行结束，FullSyn 的值则是所有可能的同义词组。

图 3 说明了查找同义词组‘甲氧苄氨嘧啶；Trimethoprim’完整同义词组的过程。图中 (1) 表示同义词组入完栈的状态，字符串变量值为同义词本身；(2) 栈顶元素‘Trimethoprim’弹出，并从原始同义词表中查到该元素的所有同义词；(3) 查询结果与字符串变量 FullSyn 比对并入栈后的状态，字符串变量值为‘甲氧苄氨嘧啶；Trimethoprim; 甲氧苄啶；甲氧苄胺嘧啶；TMP;’栈变量新入栈‘甲氧苄啶’和‘甲氧苄胺嘧啶’二元素；(4) 栈顶元素‘甲氧苄胺嘧啶’弹出，并从原始同义词表中查到该元素的所有同义词；(5) 查询结果与字符串变量 FullSyn 比对并入栈后的状态；(6) 栈空，字符串变量值为同义词组‘甲氧苄氨嘧啶；Trimethoprim’完整同义词组，查找结束。图中省略了‘甲氧苄啶’和‘甲氧苄氨嘧啶’的查找过程。查看各词来源和详情，可采用图 2 的方法。

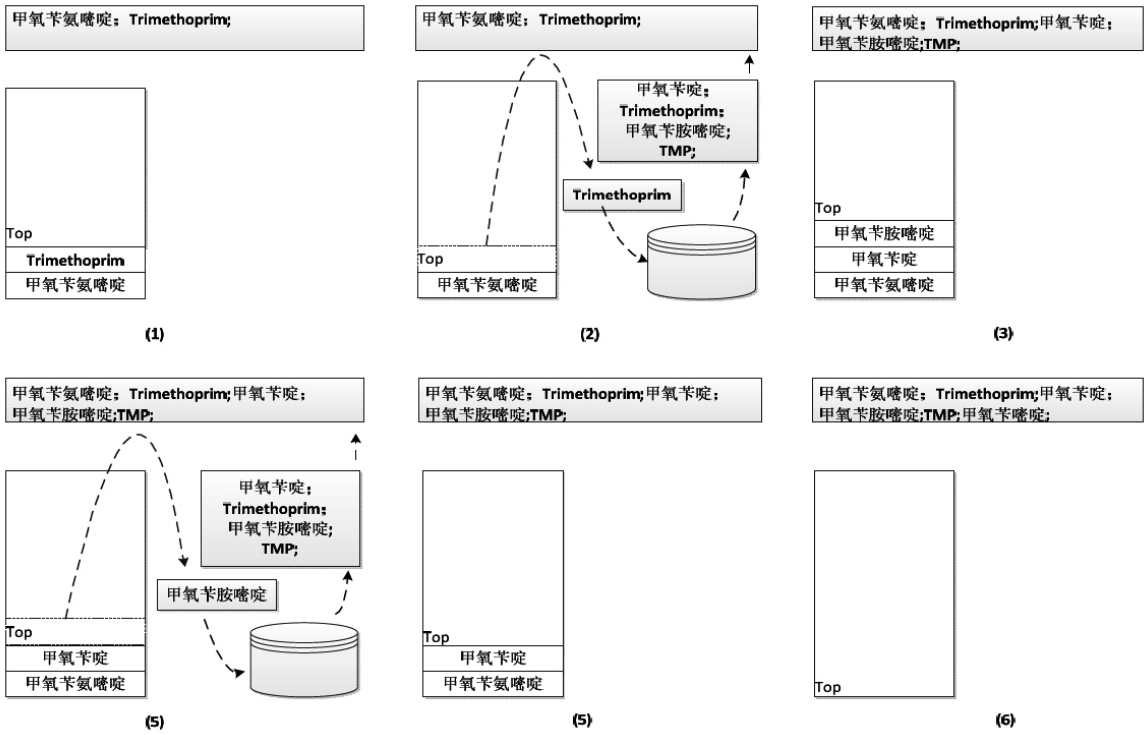


图 3 完整同义词组查找过程

三、系统设计与实现

系统架构如图 4 所示，由数据访问层、业务逻辑层、Web 服务层、用户界面层组成。数据访问层主要完成对原始同义词表和同义词基表的增、删、改、查操作。

业务逻辑层主要负责同义词组拆分，过滤，同义词自动连接算法实现，完整同义词组查找算法实现。

Web 服务层主要将各种义词查找功能进行服务化，以便进行分布式数据访问，为广泛应用同义词库奠定了技术基础，同义词服务主要提供如下功能：

判断一个词或词组是否存在同义词组；

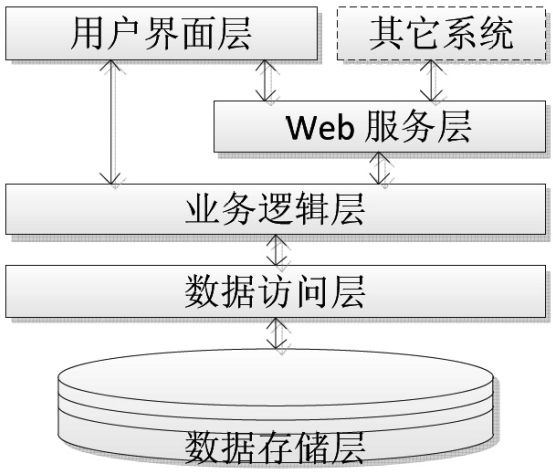


图 4 系统体系结构

获取一个词或词组经过确认为正确的同义词组，主要用于文献检索；

获取一个词或词组经过确认为正确和未确认的同义词组，主要用于数据加工中化学结构查询；

获取一个词或词组的经过确认为错误的同义词组，主要用于系统自动连接义同义词组；

获取一个词或词组的经过确认为无意义的同义词组，主要用于错误提醒和质量控制。

用户界面层主要指同义词维护界面，如图 5，实现了以下功能：同义词维护任务申请、完成确认流程管理；同义词组查询，来源信息查询；同义词组正确性确认；同义词基表增删改查；原始同义词表增删改查；同义词组不参与连接标记；重新生成同义词组，以处理计算不能处理的工作。

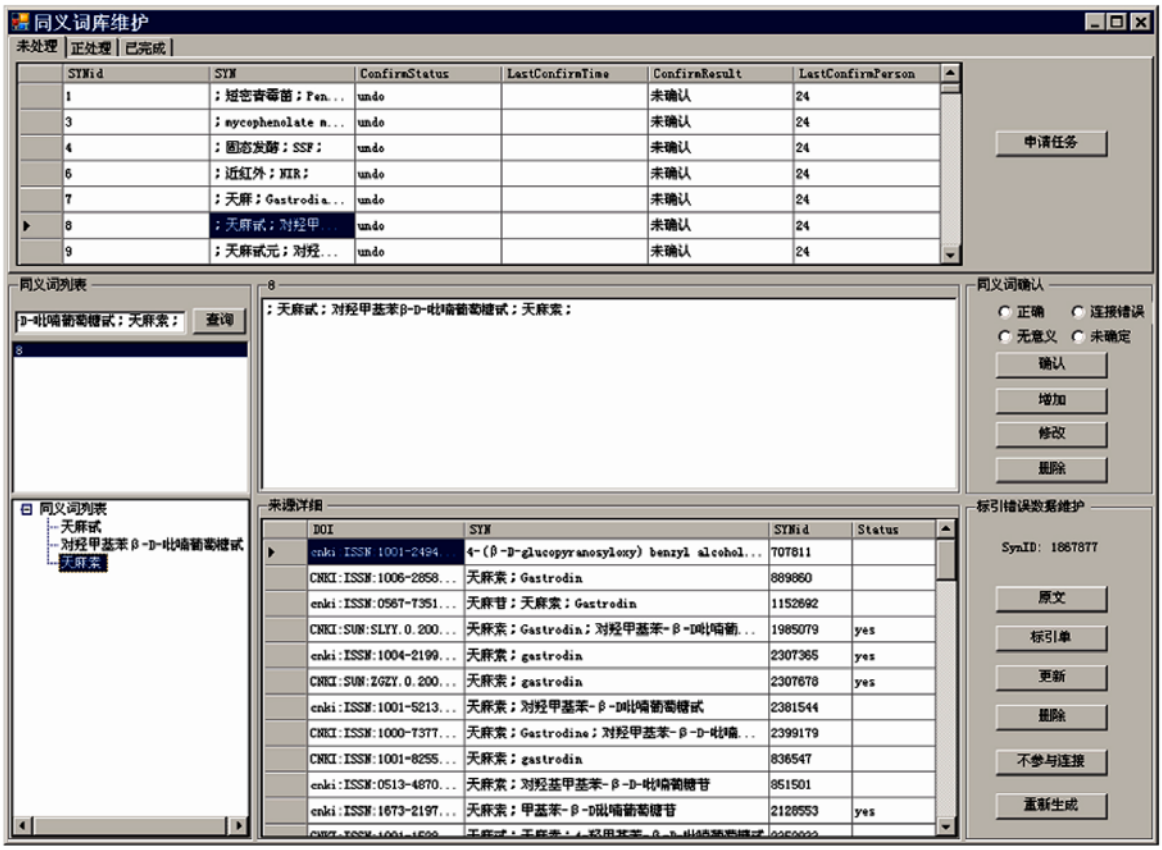


图 5 同义词维护界面找方法

在系统实现方面，数据存储用 Sql Server 2005 数据库，基于 .Net 框架，C# 语言开发，采用 Winform 用户界面。

四、同义词库应用

（一）化学结构标引

化学结构标引的基本内容为，化学结构名称、职能符、来源等信息，若化学结构库中不存在该结构，需要绘制 Mol 结构，同时提供同结构的 Gif 图片。目前，加工系统对化学结构查询还不具备结构检索功能，化学结构查询主要通过名称获得。因此，

同义词在化学结构标引中的意义极其重要，若没有同义词帮助，不同名称同一的化合物其化学结构将被多次绘制。绘制一个化学结构不仅工作量大，同时还要求较高专业技能，重复绘制不仅严重影响了加工效率，还降低了化学结构加工质量，常会出现多人绘制的同一化合物结构，结果不相同的现象。通过同义词扩展，可以大幅度降低重复绘制几率，从而显著提高加工效率和质量。在非专利数据加工系统中，提供了化学结构标引工具，如图 6 所示，以标引‘黄示灵’为例，若直接通过名称查询，系统提示没有该化学结构式，通过同义词扩展，找到了结构式。可以看到，在化学结构库中‘黄示灵’并没有直接存储其化学结构，是通过同义扩展和关联，找到同义词‘木犀草素’，‘木犀草素’在化学结构库存储有化学结构，‘黄示灵’最终引用了‘木犀草素’的化学结构，而不需要重新绘制一遍相同的化学结构。

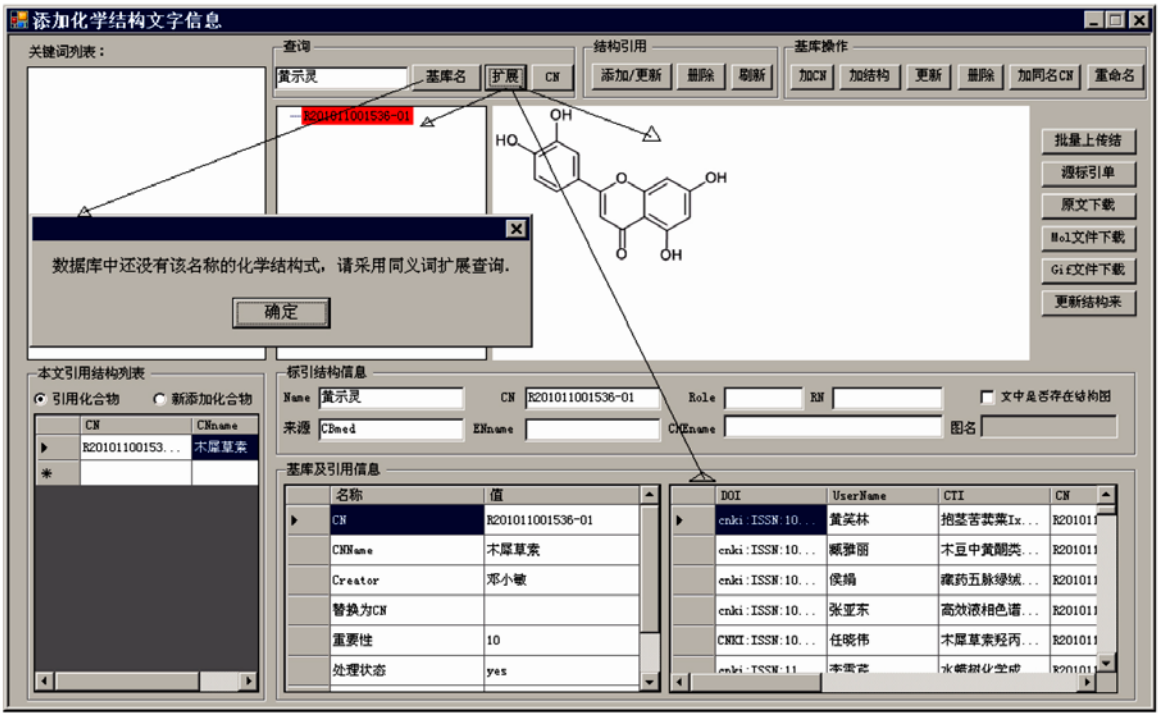


图 6 化学结构标引工具

（二）同义词半自动标引

用需要标引的同义词查询已标引的同义词组，将查询结果作适当修改和调整，自动插入到标引单，可节约输入时间，同时对同义词校验也有一定参考价值，此外，IPC、范畴分类号和方剂信息标引都可通过同义词扩展，查询已标引的历史信息，为当前标引提供参考。

（三）文献筛选

在筛选含可专利技术文献时，标引员通常擅长某一领域或知道某一关键词相关的文献可标引，用关键词搜索筛选，若用同义词扩展后进行搜索筛选，将获得更全面的结果，可以显著提高筛选效率。

（四）非专利文献检索

在文献检索中，可以广泛用于各类查询，无论是初级、高级、专业查询，还是针对题目、摘要、关键词、内容查询，只要用到词语进行检索都可以进行同义词扩展。以用关键词‘1，6二磷酸果糖’查询为例，如图 7 所示，如果仅用该词查询，命中数为 27 条，通过同义词扩展，检索到 129 条，并且其准确程度与未扩展检索时一样，具有同样的参考价值。可见同义词库在非专利文献检索中具有重要作用。

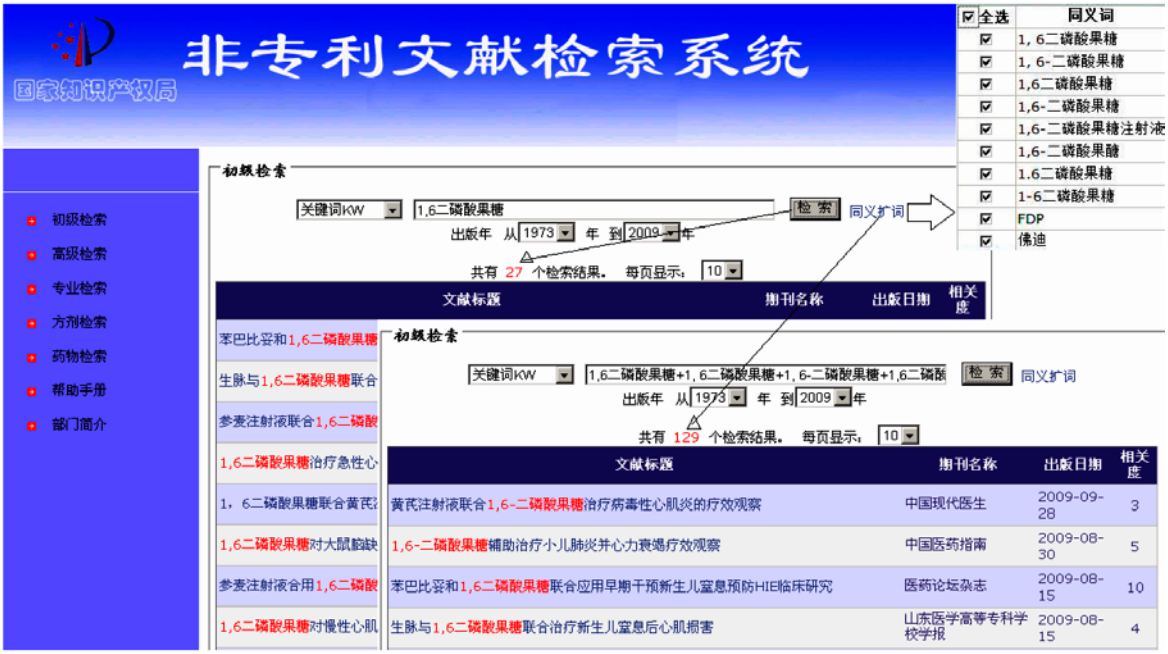


图 7 非专利文献检索

五、结论与讨论

本文通过精心设计的数据结构，同义词库中不仅存储了正确的同义词组，还存储了连接错误和无意义的同义词组，通过按需获取同义词，正确同义词组可以为正常检索服务，无意义同义词组可以提醒标引员不要再标，而连接错误同义词组可让系统选择新的连接算法，计算机不再犯同样的错误，从而具备学习能力，提高了同义词构建准确度，避免了标引员做重复工作，减少确认工作量，将构建的同义词库应用于数据加工和文献检索的各个方面，发挥重要作用，产生了积极的意义。系统用 Web Service

封装同义词服务，实现了分布式，跨平台同义词方法调用接口，即使同义词内部算法发生改变，也不需要对调用服务系统重新编译，为广泛应用同义词库奠定了技术基础。

由于同义词的情况特别复杂，目前主要同义词来源于医药化学领域，如果领域扩大，还会出现新的情况，需要改进算法。此外，当前同义词来源于标引文献，同义词的完整度受到标引期刊范围的影响，对于仅将同义词库用于所加工的非专利文献检索是没有问题的，如果把应用范围扩大，超出了所加工的非专利文献范围，应用效果可能受到一定影响。

（专利检索咨询中心　杨晓春　审校）