关于非专利文献数据深加工策略的思考

专利检索咨询中心 张秉斋

摘　要：本文就目前的非专利文献数据深加工策略概括性地提出了一些观点，涉及期刊选定、文章筛选、加工深度和精度以及关键词、摘要、IPC、化学结构标引等多个方面。

关键词：非专利文献；数据深加工策略

非专利文献数据深加工工作是我局“十一五”信息化建设的重要内容之一，目前医药类中国非专利文献数据深加工试验阶段已近结束。通过一年多的试加工实践，发现了诸多问题，特别是在加工策略方面。其中一部分问题是由于没有实施或没有完全实施原定的加工策略造成的；一部分问题是在试加工过程中新遇到的。制定、落实科学高效的数据加工策略是生产高质量、高利用价值的非专利文献数据的基本前提，因此应根据试验情况，在深入征求有关方特别是审查员意见的基础上必要时对现有加工策略做一些修正。

一、关于期刊的选定

我国现有公开发行的期刊有9000多种，科技期刊有4000多种，其中所含的文章数以千万计，不可能对所有期刊进行“深”加工，也没有必要性，因为相当一部分期刊的内容与专利审查毫无关系或关系不大。鉴于此我局选定了162种期刊作为专利审查过程中的“我国非专利最低文献量期刊”（国家知识产权局128号文件），并拟对这些期刊中的数据进行深加工。在试加工阶段仅对其中的7种期刊进行深加工，以在获得加工经验和对试加工数据进行评估的基础上，再稳步推进162种期刊的数据深加工，见图1。

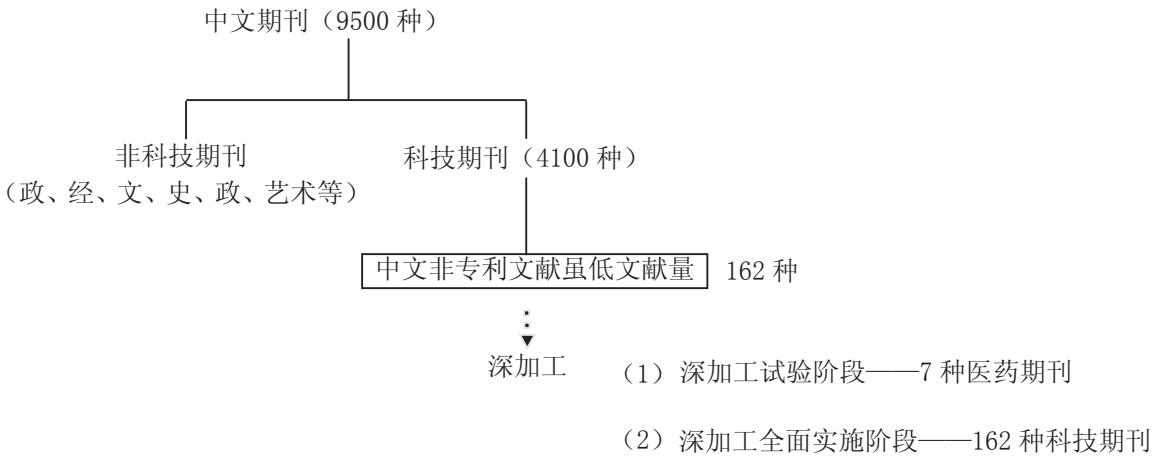

图 1 我国非专利最低文献量期刊的选定

我国非专利最低文献量期刊列表原则上是动态的，应根据审查实践即检索报告中期刊的引用频次做必要的调整。在对非专利最低文献量期刊列表做调整时，应将检索报告中期刊的引用频次和各个领域的专利审查员的意见作为选刊的第一因素，而不应过多地考虑期刊是不是“核心期刊”,是不是“双高”、“双优”和“双百”期刊，以及所谓的国际影响力。有的核心期刊主要涉及科学理论、基础研究方面的内容，不提供技术方案或技术教导（technical teaching），它们在专利审查过程中被引用的频次很低，但也被选入我国非专利最低文献量期刊列表，例如《力学学报》，在 2004-2006 年审查年度中被引用的次数仅为 1 次。在这 162 种期刊当中有 17 种英文版期刊，它们在专利审查过程中被引用的频次更低,有的引用频次甚至为零（按 2004-2006 年审查年度统计），例如《中国海洋湖沼学报（英文版）》、《理论物理通讯（英文版）》等。

另外，在对我国非专利最低文献量期刊列表做调整时，似乎应侧重考虑那些没有被国外知名数据库收录的期刊，而不是重复收录、加工，只有这样才能做到与国外知名数据库互补，在与其他局进行数据交换才更能凸显价值。

二、关于加工深度和精度

欧洲专利局（EPO）认为，数据应具备正确性、及时性、完整性。因此 EPO 在非专利资源建设方面在保证数据正确性的前提下把重点放在非专利文献的快速、大量收集方面，以实现数据的及时性和完整性。EPO 通过数据格式标准化，将所收集到的大量非专利数据与专利数据一并纳入到其内部检索系统 EPOQUE 中。2008 年，EPOQUE 中装载有 120 个

内部数据库，其中有专利数据库 55 个、非专利数据库 32 个，总共包括约 4.4 亿万条可检索记录 [1]。从图 2 可以看出，尽管 EPOQUE 中的非专利数据库的数量少于专利数据库的数量，但是非专利数据记录数从 2006 年开始已经超过了专利数据记录数，而且近几年非专利数据记录数的增长速度要高于专利数据记录数的增长速度。

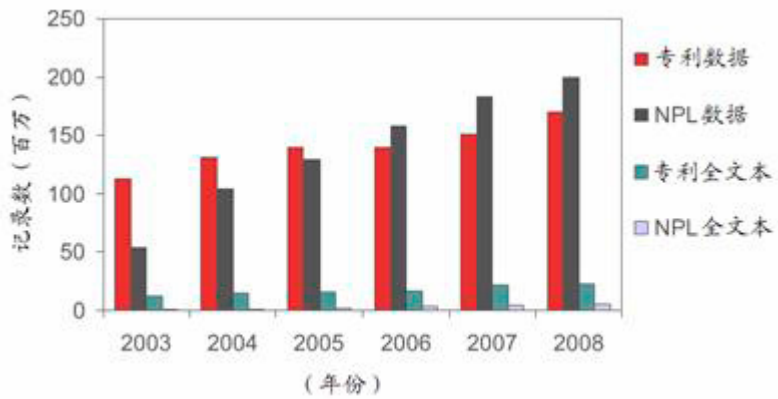

图 2 EPOQUE 中非专利与专利数据的记录数 [1]

在 EPOQUE 收录的非专利数据中，EPO 仅对很少量的重要非专利文献（主要是检索报告中引用过的非专利文献以及 PCT 最低文献量列表中所列期刊中的、审查员认为在专利审查中很可能被引用的文章）做加工，加工深度仅限于给予非专利文献号（XP 之后接着 9 位数字）和 ECLA（目前仅 10 万左右的记录具有 ECLA），而不涉及其他项目。以下是 EPO 的 NPL file 中的一条记录 [2]。

1/1 NPL - (C) EPO

AN - XP000678661

TI - ADAPTIVE MODEL-BASED HYBRID CONTROL OF GEOMETRICALLY CONSTRAINED ROBOT ARMS.

AU - WHITCOMB L. L.; ET AL

DT - J (Journal Article)

SO - IEEE TRANSACTIONS ON ROBOTICS AND AUTOMATION, 1997, vol. 13, no. 1, page(s) 105-116

JN - IEEE TRANSACTIONS ON ROBOTICS AND AUTOMATION

NU - ISSN 1042-296X

PD - 1997-02-01

EC - B25J-009/16L1

P - 2003-01

而我局的非专利数据进行深加工项目涉及的标引条目较多，包括标题、摘要、关键词、IPC、化学结构等，要求的标引深度和精度也很高。加工深度、精度直接影响加工效率，如果过于追求加工深度、精度，在人力资源一定的条件下势必影响加工效率，从而影响加工后的数据纳入检索系统的进程，进而影响审查员的使用和对加工数据的反馈评价，反过来又影响加工策略的及时修正。EPO 的非专利文献加工策略值得借鉴。

三、关于期刊中的文章的筛选

在非专利文献数据现行深加工流程中（见图 3）可以看到有一个非常重要的筛选文章的步骤。但在实际操作中只是除去了明显不属于科技类的文章，其余的基本上都保留下来并予以深加工。合理的策略似乎是对科技类文章做进一步的筛选，以选出那些在专利审查中被引用概率高的文章，即那些包含可专利的技术方案的文章，以及研究结果对于专利审查有引用价值的文章（例如药理研究类文章）。应剔除涉及不可专利主题的文章，例如关于手术方法、借助仪器诊断疾病的方法等类型的文章，纯基础性科学研究文章，例如植物的种原分析、植物的染色体数目分析、地质演化等类型的文章。

目前我局非专利文献数据深加工策略中缺少一个筛选文章的标准或规则，笔者认为可以根据专利法第二十五条来制定筛选文章的基本标准。对于常规技术领域，也可以考虑以 IPC 作为筛选文章的间接标准，即如果文章的主题或其中所包含的技术方案在 IPC 中具有分类位置，则选定，如果没有分类位置，则去除。

文章筛选不当，不仅会影响加工效率，浪费加工能力，将来还会增加检索的系统性噪音。

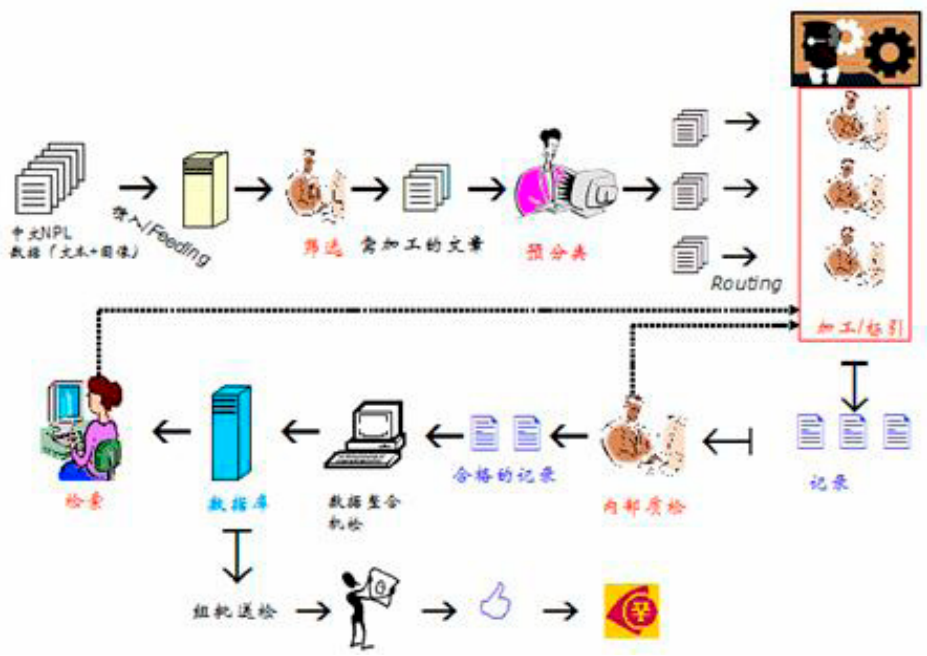

图 3 非专利文献数据深加工流程

四、关于标题、摘要、关键词

随着计算机能力(运算速度、存储量)提高和各种算法的发展，文献的储存和检索环境已经由传统的纸介质转向电子介质，检索手段也由手工检索转向计算机检索。因此，数据深加工的策略也应“与时俱进”，做出相应的调整。

在电子文献环境下标题、摘要具有检索和浏览两种功能，如果数据库中具有原文（图像文件或全文本文件）资源可以链接的话，那么标题、摘要的浏览功能就相对弱化。

从关键词的原始定位来说，其功能是专门为了检索，因此，如果数据库在提供了人工深度标引的关键词的情况下，标题和摘要的检索功能也就自然弱化了。

也就是说，如果提供了人工深度标引的关键词，数据库又能提供与原文的电子链接，那么标题和摘要的加工可以相对简化一些。

就标题、摘要、人工深度标引的关键词集合当中所含的关键检索要素的数量而言，一般情况下，关键词集合≥摘要≥标题；就检索时可能产生的噪音而言，一般情况下，关键词集合≤标题≤摘要。

摘要一般可以分为两种类型，一种是指示性（indicative）摘要，另一种是信息性（informative）摘要。目前在我局的非专利数据加工实践中所采用的摘要标引策略倾向于提供具体细节，属于后者。在高度方便的电子文献检索环境中，信息性摘要的重要性已经降低。至少对于一部分非专利文献，例如综述类文献，应当根据情况允许使用指示性摘要，以简化摘要的标引。

五、关于 IPC

使用 IPC 对非专利文献进行分类对于我局来说是一个新生事物，在实践中遇到很多问题，亟待解决。例如，关于用 HPLC 分析药用植物中的化学成分的文章非常多，在加工实践中一般都给了 G01N30/02，这似乎不太合适，值得进一步商榷，因为文章中对 HPLC 系统本身没有做出任何改进的描述，只是使用现成的 HPLC 仪来分析了一下化学成分而已。又例如，关于用显微镜通过观察药用的组织结构来判断是否是道地药材的文章，给予关于显微镜本身的分类号显然也不合适。

尽管 WIPO 在其 IPC 发展战略中鼓励各局将 IPC 应用于非专利文献，但并不意味着所有科技文章都适合用 IPC 进行分类。应该对不同主题的文章区别处理，对于适合用 IPC 分类的，则标引 IPC 分类号；对于不适合的，则不标引 IPC 分类。“宁缺毋滥”是合适的策略。若在非专利文献数据深加工中牵强附会地标引 IPC，势必将失去 IPC 在检索中的应有的效力，也就失去了标引 IPC 的意义。

六、关于化学结构

合理的策略是先购买或开发化学结构检索系统，然后按照系统的具体要求，规范地画出化学结构，否则即使都是可检索的mol文件格式，也不能保证所画的化学结构都符合化学结构检索系统的要求。例如，法国工业产权局的化学结构标引规则对化学结构中的异常原子量、异常化合价、电荷、配位键、均化键、互变异构、糖、盐等的画法都有明确具体的规定，以与其化学结构搜索引擎相配套。如果没有配置化学结构检索系统，对于一种化学结构而具有多种不同画法的化学结构就不能提出一个明确标引方式，因此必然会遗留一些问题。

参考文献

1.EPOQUE- the EPO system for finding relevant prior art, Claus Albrecht, 2009

2. NPL File - layout Description, EPO, April 2005

（专利检索咨询中心 杨晓春 审校）