一个经济实用的全文资料信息系统的建设

北京信息技术应用研究所 解立平

一、问题的提出

改革开放以来，国内各部门都把尽快改变自身资料工作环境的任务列入议事日程，引进了较大型计算机系统。以适应快速发展的科技社会和信息化时代。但在普及推广计算机应用工作中，几乎同时遇到同一个难题：即如何将原有的和现在的大量资料输入到计算机数据库中。一些专业部门仍延用了六、七十年代的传统资料工作方法，即由专业标引人员（包括专职标引队伍或业余标引人员）对资料进行人工标引，再把规范化的数据装入数据库。这对某些具有特定专业方向、并已具有一定经验和规模的标引队伍的部门，从事对一定量的资料进行数据规范化处理和转换是可行的。而对于以前没有专业标引队伍的单位来讲，要想把今天接近于爆炸式的文本信息资料通过六、七十年代的标引方法进行整理，再送入数据库，在人力和财力上是相当困难的。由此就产生了这种现象，计算机技术部门用较快速度建立了数据库，而用户部门没有能力来保证整理、标引出资料数据，或者没有力量再去更新、补充数据库中数据内容，致使建立的数据库因无实用的内容而无法发挥效益。如何打破僵局，尽快用现代化手段改善业务人员的工作环境、提高他们的效率是我们计算机技术部门在新势下面临的重要和需要迫切解决的任务。

能不能产生一种新的模式？它由计算机技术部门依靠自身力量，利用当今高科技技术，将用户急需的印刷文本或其它形式的资料放入计算机，直接提供给用户使用，满足他们需求。由此打开应用的局面。

二、利用高科技技术打开突破口

要依靠计算机技术部门自身力量来找到突破口，我们认为技术部门首先要解决以下的问题：

1.要避开标引工作的限制

由技术部门组织资料入库，首先要解决技术部门自身无标引力量和经验的问题。既使技术部门可以组织标引，但能否根据所有业务用户的业务观点来做文摘及主题标引，并被所有业务用户所接受，也是一个问题。所以要采用避开对资料进行标引的局限性。

2.要保证数据转换的速度及连续性

对于大量的信息资料，如何能利用高科技技术将数据转换的效率提高，以适于用户对当今大量资料的需求。同时这种转换的效率还要保证以最小的代价和可能的机构允许情况下去实现。

3.绝对保证所有入库资料数据的安全性

当大量资料入库以后，将使用户人员的工作环境得以改善。入库的大量资料成为支持用户的基础。用户可以不再保存大量的原文资料。数据库的可靠性和安全性也就成为至关重要的因素。要采取有效措施来保证数据的安全。

八十年代中后期，全文本方式的数据库在国际和国内出现并展现出很强的生命力。这种全文本方式数据库利用了当今计算机硬件存储介质价格越来越低，CPU 运行速度越来越快的优势，一改传统数据库需将资料进行相应的规范化标引处理后再入库方式，将全文本资料装入数据库。数据库软件将全文数据中的每个词或字做相应的切分、抽词排序等处理。用户只要根据所需查资料的相关查询词并结合逻辑关系，通过不断地筛选，就可直接找到所需的全文文本资料。这种新型技术不但避免了大量需要标引人员对资料进行认真、细致入库处理的限制，加快了资料入库的速度，也克服了由于标引人员因经验、方法、观点不同以及研究方向、角度的不同而产生的标引思路及尝试不一致性，增大了数据库内资料的共享

程度。为资料工作的计算机现代化普及找到了一种新的应用方式。另外，在社会进入了信息时代后，对于我国来讲，大量的汉字资料的输入问题变得十分突出。新出现的汉字计算机自动识别技术为自动转换汉字文本资料提供了高效的转化方式。这种新技术改变了长期以来靠手工输入汉字的落后方法，利用计算机的高效的特点将大量印刷文本资料自动转换成计算机化的数据。这种转化方式大大减轻了数据录入人员的劳动强度。也大大降低了录入成本。为适应当今大量文本资料的汉字转换打下了基础。

三、系统软硬件构成环境

根据以上高科技的发展情况，我们设计出了一个包括三个子系统的全文信息系统，以图利用我们自己的力量来建设一个方便于用户的资料数据库系统。

1.硬件环境

IBM 4381　中型计算机

386 微机　两台

HP 9195　扫描仪两台

5550 微机仿真终端若干

2.软件环境

IBM　VM／SP 操作系统

浙江省经济信息中心开发的《通用全文检索软件 H-CGRS》清华大学电子工程系图像教研室开发的 OCR 汉字识别软件

四、汉字转换子系统

汉字转换子系统的功能是将非数据化的燃料转换成计算机的内码。转换工作是由计算机技术部门组织的录入修改人员承担。由于我单位用户大量需用印刷文本资料，所以建立以印刷文资料为主要内容的全文资料库首先要解决的任务就是把要入库的各种印刷文本形式的资料进行汉字代码转换。而对这个汉字转换子系统的要求是：

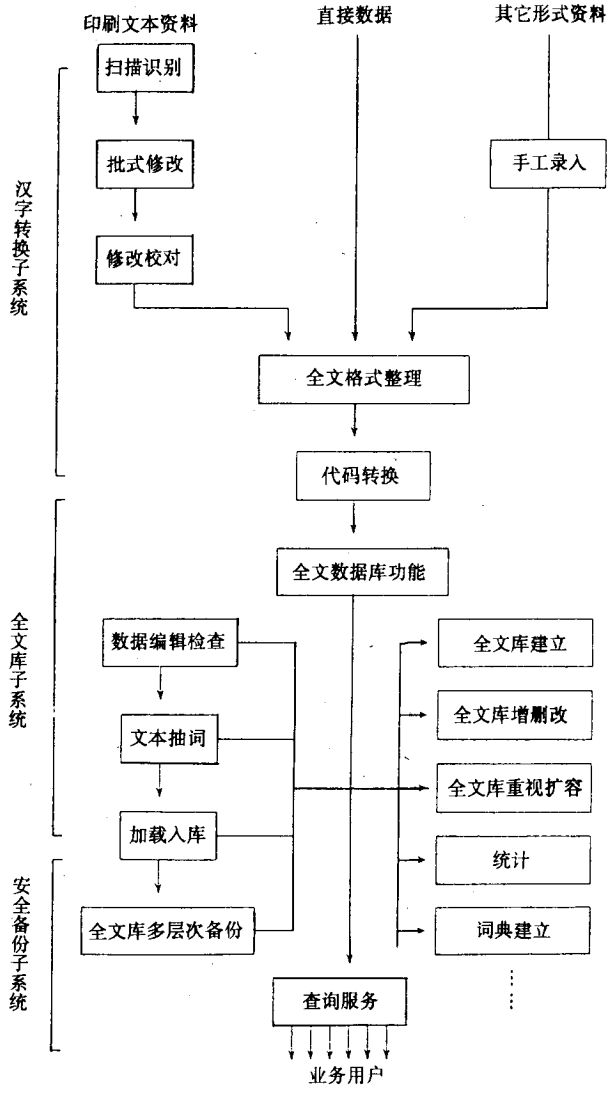
保证资料转换的准确性；

转换的成本代价要尽量小；

转换的速度要能适应所建立的全文库加载的数据需求。

对于一些非印刷形式的资料也应可以组织录入人员进行转换。

系统工作流程图如下：



为了使系统达到最低的成本、最快速度要求，在新系统中必须采用新出现的计算机汉字扫描识别技术。利用它一方面提高转换的准确性，减少录入中的丢段漏字，一方面降低了转换过程中人工修改的费用。我们选用了清华大学电子工程系图像教研室开发的汉字识别 OCR 软件。此 OCR 软件采用了对印刷文本中的黑体、宋体、楷

体和仿宋体汉字同时进行识别的方法。提高了汉字转换效率。而且对我们所要进行转换的印刷比较差的特定文本资料，也具有较高的识别率。用户界面也十分利于用户学习和使用。我们将此 OCR 软件作为汉字转换子系统的核心。形成一个适合于各种形式数据转换的运行机制。录入修改人员利用 OCR 软件将大量印刷文本逐而扫描并识别成计算机数据文件。而后由录入修改人员将数据文件与原文资料进行对比修改。完成校对工作。保证转化数据的正确性。此汉字数据转换子系统能适应数据的、非数据的、印刷的和非印刷的各种形式资料的数据转换。两年来，技术部门的管理和技术人员通过对子系统内的录入人员的管理，保证了此子系统在低费用、高效率状态下正常运转。并保证每天 10 万字左右的入库数据资源。

为了提高转换速度，技术人员在微机上编制了批执行程序，将光扫识别软件的工作方式由每次单页扫描识别变成先将所有文本资料扫描成图像文件，而后统一进行批作业方式的连续自动识别。避免了过多的人工干预。提高了转换效率。

在经过一段运转后，我们发现该识别软件对我部入库的文本资料具有固定的误识规律。为了进一步加快汉字转换速度，技术人员在 IBM4381 机 VM / SP 系统下用 REXX 宏语言编制了对有规律的错识字进行批作业方式修改的程序。它可以在录入修改人员校对修改之前，利用计算机进行第一次的自动修改。将有规律的固定误识错字进行修改替换。既减轻了修改人员的工作强度，又加快了资料的转换速度。取得较好的效果。

整个数据转换子系统在近两年的运转中，用扫描识别软件转换了近 5000 万汉字资料。用其它方式转换了近 1000 万汉字。用比外界录入费低得多的代价保证了全文库的加载数据。(每万字录入费用仅 3 元人民币)。基本达到了转换子系统的预期目标。

五、全文数据库子系统

全文数据库子系统的管理、维护工作主要由技术人员负责。它是整个全文信息系统的关键。它的功能必须能满足以下要求：

- 全文数据库软件；
- 全文方式的查询方法；
- 各种全文库的管理和维护功能；
- 必须可以在 IBM4381 机上的操作系统下运行；

通过了解和调研，我们找到了国内可以在 IBM4381 机上运行的以全文方式工作的全文软件-浙江经济信息中心开发的《通用全文检索软件 H-CGRS》软件。该软件虽然暂不是标准的全文方式的数据库系统软件，但通过详细的调研、分析和试验。认为该软件通过再增加一些适应全文工作环境的系统功能，就可能基本达到全文数据库的功能，达到我们对全文信息系统建设的要求。此软件是工作在以对字典词汇进行切分基础上的。软件根据全文库中的词典对入库数据中的词汇进行切分，并建立这些词汇的倒排索引文件，用户通过词典中存在的词汇进行查询。如果用户希望用抽词词典中没有的任意字或词去查询资料，需首先进行对词典词汇或其它库可查询字段的一次查询，而后再对一次查询结果进行二次串匹配查询。虽然这将给用户查询增加不方便，但作为能在 IBM 4381 机上工作的全文软件，我们认为它仍然具有较灵活的建库方式和较完善的库管理功能。只要再增加一些适合于全文工作环境的功能，是能适应我单位对大部分汉字全文资料的建库需求。所以我们把此软件作为全文数据库子系统的运行核心。此全文数据库子系统通过技术人员的管理主要完成以下工作：

- 将用户对资料使用的需求用软件命令方式描述成库内的字段、类型、大小等参数。并生成全文数据库。
- 管理各个全文数据库的用户使用权限。
- 建立所需的全文库抽词词典。
- 对加载数据进行必要的自动检查和数据的入库格式整理。
- 对加载的数据进行抽词。
- 数据加载。
- 全文库内数据的各种统计。
- 全文库的备份。
- 全文库内文献的删、改。
- 全文查询功能。
- 其它一些用户所需的功能。

由于所建立的全文信息系统具有全文库数量多，单个库容大，并且加载频繁的特点，为了进一步改善全文系统的整体性能的扩展性，结合我单位的实际情况，我们会

同全文软件开发单位着重对全文软件增加了一些改善系统整体性能的措施。对于原全文软件没有的而全文信息系统又需要的功能，我们与原开发单位共同研究、实验，加以解决。使系统的整体性能达到全文信息系统所需的运转环境。我们做了以下技术改进工作：

·结合用户人员计算机知识水平不高的实际情况，技术人员与开发单位配合在 VM／SP 系统中用 REXX 宏语言编制了适合于业务人员使用的每个全文库对话屏幕界面。使用户人员可以简易、方便地将所要查询内容的字段项、查询词及有关信息等数值直接放入计算机屏幕。由计算机程序自动为用户生成内部标准的逻辑查询表达式。不用业务人员自己去组成复杂的表达式。减少了差错。全文库中按以及组合查询等方式中均提供了用户只需键入相应字段值的查询屏幕。内部软件对于用户提供的字段值自动生成查询表达式。极大地方便了用户的使用。受到欢迎。

·根据用户人员的研究范围并结合所需要入库的文本资料内容，技术人员对我单位的主题词表内的 9000 多条主题词进行了筛选。把一些常用的人名、地名等业务人员可能查询的词汇与所选择出的主题词共同生成适合业务人员查询的抽词词典。

·全文查询软件原来只能为系统内全语文库提供一个抽词词典。这对于将要建立的大量不同类型全文库是不相适应的。某些全文库可以共用一个抽词词典。但有些全文库则需要不同类型的或与全文库对应的库词典。通过与原开发单位进行技术上的分析后，在原来软件上增加了建立了多个全文库词典的灵活功能。增强了实用性。

·全文查询软件原来只能对一个工作数据盘进行建库。我们发现对于我单位所建容量较大的全文库，单一工作数据盘只能建立少量全文库。这无疑是对全文信息系统发展的很大限制。我们配合全文库软件开发人员增加了全文软件的新功能。使得在 IBM4381 机全文软件主用户的任何虚盘上都可以建立全文库。使此全文信息系统的可扩展性满足了我单位的需求。

·原全文软件的库工作方式是首先根据全文库最终可能的最大数据量预先建立起全文数据库容积。以后不断向库里输入数据，直至装满。这种方式不适合我单位每天频繁地追加数据使库容量不断增长的使用情况。我们配合开发单位在全文库软件上增加了所需的新功能。当没有库空间的情况下，能将全文库的空间按任意比例增大。实现了使全文库由小到大不断扩展的需求。

·原来全文软件的库备份功能只能对容量可以压缩在一盘磁带上的全文库进行备份。而对于我单位较大全文库则无法实现跨带备份功能。

两年来，我们从完善我单位信息系统以适用于用户应用环境的角度出发，不断地与开发单位合作。共同在增强此信息系统的整体功能上做了大量工作。使全文信息系统的各种功能基本满足了我们的建库要求。两年来，建立了 6 个全文库。数据量达 6000 万汉字左右，各种系统功能正常。初步地发挥了系统效益。

六、安全备份子系统

全文信息系统中全文库数据资源的安全是十分重要的。当各个全文库不断增大，而且对用户的实际工作起到支持作用时此系统的安全可靠性成为此系统成败的至关重要的因素。为此我们仔细地分析了全文信息系统内各个环节的可靠性，并结合在日常全文库运行工作中遇到的问题在系统整个流程容易出问题的环节上采取了相应的安全保护措施。完成了从原数据、全文库的盘和带到 IBM 系统级的四层次备份。确保了全文系统运行的可靠性。

安全备份子系统的工作主要由全文库的技术负责人员和操作系统管理人员负责。我们采取以下四层次的备份措施，以保证全文数据和数据库的完整可靠：

1.源数据的可靠备份

全文资料信息系统是由多个全文子库组成。对于每一个全文子库的源数据备份是在此库数据当天加载之前，将最后验证正确的源数据和抽词后生成的标准加载数据分别做一次文件备份。经过一段时间后，再将积累起来有一定量的源数据做系统的磁带备份。以用于今后数据的交换及必要时候的数据重新入库工作。保证最基础数据的万无一失。

2.每个全文子库主要索引文件的盘备份

一定时间后，对每个全文子库的主要索引倒排文件进行统计操作，验证库统计结果正确后，用编制的批备份程序将此全文子库的 4 个主要的文本文件和索引文件备

份到另一个盘中。一旦加载失败损坏了库文件，就可利用此备份库的文件去恢复工作数据盘上的全文子库主要文件。这种方法避免了每次加载时的备份过程。加快了数据库加载速度。

3.每个全文子库的磁带备份

利用全文信息系统软件自身的备份功能，定期将每个全文子库备份到磁带上。一旦发生工作数据盘及备份盘均损坏的情况，即可用此全文子库备份带对全文库进行恢复。

4.系统磁盘的整体备份

这是利用操作系统的系统备份功能由系统管理员或操作员为保证系统安全而对每个磁盘体做的备份。定期由全文信息系统的负责人将属于此信息系统的磁盘体进行备份。这种备份包括对此全文信息系统工作盘的备份。在万一发生较大的不可预见性的全文库损坏时，做整个全文信息系统工作模块和各个全文库的数据，文件的恢复。

通过以上四层次的备份措施，基本保证了全文资料数据的可靠性，保证了全文资料信息系统资源的安全。经过全文信息系统近两年的运行以及几次出现的问题，证明了我们所采取的备份措施是可靠和可行的。

七、结束语

在两年的建设全文资料信息系统过程中，我们按预期的设想走出了一条由技术部门领先自身的力量建设全文资料信息系统的新路。找到了一种能使用户部门直接享用计算机资源的计算机推广普及模式。现建立的6000多万汉字容量的全文数据资料库已向用户提供使用。也得到了用户积极的反馈信息。为进一步搞好计算机的应用打下了基础。今后，我们将加快系统内全文子库的建设工作。特别是对用户共用量大并且急需的资料，放在优先的地位。更好地提高系统的效益。为计算机应用做出更大贡献。