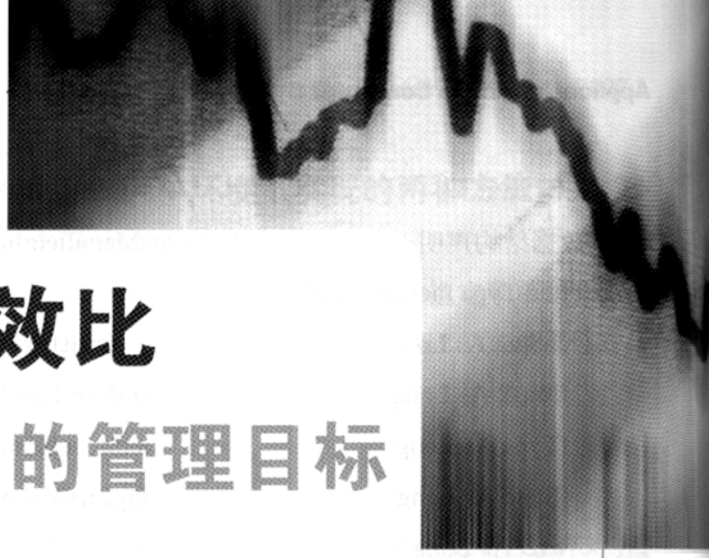

降低系统全寿命费效比
——全文检索系统应用的管理目标

郑 宏 (首都经贸大学信息学院 100000)

摘 要 本文通过分析全文检索系统在机关档案部门的应用特点，提出将降低系统全寿命费效比作为全文系统的管理目标，寻求最佳系统全寿命费效比，并提出了一个将高速扫描、OCR与全文系统结合的应用方案。

关键词 全文检索 费效比

1 引言

全文检索系统现已成为许多机关档案部门管理信息系统的标准配置。但部分单位的管理信息系统投资巨大却不能按照期望发挥管理作用，原因之一就是这些系统中没有大量的、有意义的信息，不能对管理使用人员产生吸引力。造成这种情况的原因并非这些单位没有大量信息，而是由于在系统的设计、建设时期对机关档案部门全文检索系统的应用特点和系统全寿命费效比的认识、规划不足造成的。

全文检索系统在机关档案部门的应用特点与图书情报、出版部门有很大的不同。首先，全文检索系统在办公系统中虽然只是一个子系统，但具有相对的独立性，表现在全文数据的流动具有单向、不可逆性，一般只从办公系统的其他子系统向全文系统输入，而全文系统一般不向其他子系统输出数据，全文数据管理上可以相对独立；其次，运行维护费用要求低，机关的办公系统主要服务于内部，没有盈利的需求，机关费用预算严格，全文系统运行维护费用必须维持在低水平上；再次，操作必须简便，一般的数据输入和查询一、两步操作就可以解决，多不过三步，使机关干部等非专业人员的培训工作降到最低。

根据以上特点，目前传统的单纯文档型全文检索系统已不能适应机关办公自动化需求。传统全文系统的数据来源一般是手工录入、原系统的数据转换和迁移。这两种方法的数据输入直接或间接地来自于手工录入，数据输入量非常小，即使1000万字信息也仅相当于50本普通书的容量，机关的实际信息量远大于此数，但为了这点信息的手工操作步骤比较繁琐，必须经过校对环节，费用高昂。迁移的数据基本上是关系数据库的结构化信息，全文数据的优越性难以发挥。由于数据量很小，全文系统的实际价值也就不大，即使原来选型时考虑到全文系统可以支持上亿字的数据，但录入、校对等费用却制约了全文系统功能的发挥，产生系统“有劲使不上”的结果。

2 系统全寿命费效比

为解决以上问题，在设计、建设全文系统时必须考虑全文系统的全寿命费用和全寿命效用，降低全文检索系统的全寿命费效比。这个问题在其他办公系统中也存在，只是在全文系统中更加突出。

系统的全寿命费效比是系统的全寿命费用与系统的全寿命效用之比，使全文系统的全寿命费效比最低而不是系统购买费用最低，是系统建设者应追求的管理目标之一。所谓系统全

寿命费用不仅包括系统的初期一次购买费用，也包括系统在寿命期内的运行、维护、管理、更新费用。许多办公系统建设者的误区是仅仅考虑了包括软件、硬件、网络等初期的较低费用而忽视了全文系统在寿命期的其他费用，使得具有很强功能的系统不能发挥作用。所谓系统全寿命效用是指系统在寿命期内发挥管理作用的总和。任何一个管理系统都有它的寿命期，随着信息技术发展越来越快，办公系统的寿命期越来越短，因此必须在较短的时间尽快发挥作用。传统的文档全文系统由于输入数据速度和费用的制约，需要一个很长时期积累数据而无法发挥作用，寿命期内的效用总和很低。因此，解决全文系统的数据高速、低费用输入问题是全文系统在机关发挥作用的关键。

系统全寿命费效比与性能价格比概念类似但有很大区别，其一，后者是静态的、一次的，前者是动态的、长期的；其二，后者是单机的、孤立地看系统本身，前者是项目的、集成的，更关注整个项目的得失。因此，系统全寿命费效比指标在管理层次上更高，关注该指标较性价比能更好地解决全文系统的应用问题。

3 集成全文检索系统应用方案

本文推荐一种全文检索系统在机关的新集成应用方式可以较好地满足系统全寿命费效比要求，即：高速扫描+OCR+全文检索，这三项技术都是比较成熟的技术，通过集成应用，使得机关原始文档无需录入校对，而通过高速扫描仪扫描输入、压缩存储备份，经OCR识别后输入全文检索系统，用户仍使用传统的全文检索方式，查看检索结果既可以是OCR后的文档，也可以调阅扫描后的电子影像文档，从而保证查询正确率。

虽然OCR不可能达到100%的准确率，但对印刷体文字的识别率一般可以达到98%以上，只要所检索的文字有部分命中，全文系统即可调阅原始文档的电子影像文件，从而保证查询者所查看内容是无误的。该模式的系统流程如图1所示：

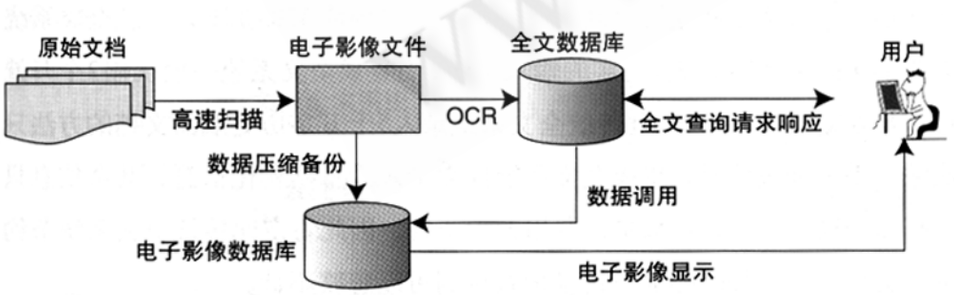

图 1 集成全文系统数据流程图

从以上系统简图可以看出，集成的全文系统充分发挥了目前成熟技术的优点并加以组合。该模式的特点是在传统文档全文系统前端增加了高速扫描仪和集成在全文系统内的OCR软件、自动入库模块，完成高速扫描和数据自动入库，后端在必要情况下可以增加备份存储系统。在前端，目前高速扫描仪的扫描速度可达A4幅面每分钟20——80余页，完全可以满足扫描速度要求，有的还支持A3幅面扫描，较大的报表图文也可用扫描方式输入系统，输入、识别的过程人工干预少、操作简单、费用低廉。在后端，由于高速扫描处理使得数据量大大增加，在必要时应将数据备份到光盘镜像服务器或磁带系统上，光盘镜像服务器和磁带机的投资并不大，一般情况下可以在办公大系统内解决无需额外增加开支。除了集成扫描、OCR和自动入库模块，传统的全文检索软件还需要增加支持关系数据库和备份数据管理的能力。

由于采用高速输入设备，数据输入量大大增加，系统无需经过一个较

长的数据建设期，在刚刚建成时就可以发挥重大作用，寿命期内的效用大大增加，运行维护费用的降低不但弥补设备投入而且有余。因此，该应用模式从系统全寿命费效比看是十分优秀的。从管理上看，全文信息的单向性使机关信息中心完全可以承担扫描输入和管理工作，由计算机中心转变成数据处理中心。

4 案例分析

以下采用系统全寿命费效比理论对全文系统的规划问题作简要分析。系统期间划分如图2所示，主要分为系统建设期和系统寿命期。系统寿命期应从软硬系统建成调试完毕之日开始计算，但系统的全寿命效用只能从系统建成并具备最低要求的全文数据量以后，即系统正式运行时才能开始计算，因此系统寿命期分为信息输入期和系统运行期两个部分，系统效用仅在运行期产生，效用是一个比较复杂的经济学概念，为简单计算，我们在评价系统效用时用全文字数作为量化依据。

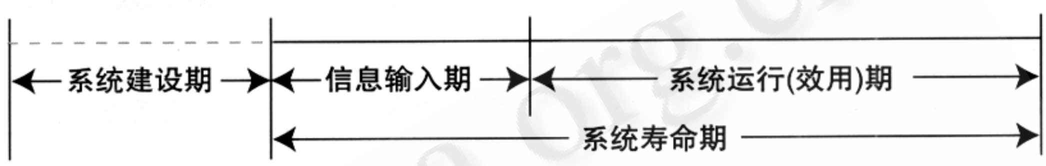

图 2　系统期间划分

评价一个全文系统的全寿命费效比有基于相同寿命费用比较系统寿命期效用高低（即寿命期内输入系统的全文数据量大小），或者基于相同寿命期效用比较系统寿命费用高低两种方式。在实践中，由于机关/事业单位的预算费用严格，我们采取前一方法评价全文系统的选型，即在全文系统费用的预算硬约束条件下，按照系统寿命期内能够输入的全文数据量大小来确定系统选型。

假设某单位需要将海量历史文档以及今后产生和接收的文档都纳入一个全文系统进行管理，我们依据以上系统期间划分作方案比较分析。

在系统建设期，投资一个传统全文系统的投资约为20万元（含网络检索软件、服务器等硬件），采用集成全文系统方案以后，需要增加影像扫描处理、OCR 处理等软件模块和高速扫描仪，软硬件投资需要增加约 5 万元。

在信息输入期，如果使该单位内部的全文系统开始正常运行并发挥应有的作用，至少应具备1亿字以上的全文数据量，即相当于一个拥有500本图书的小型专业资料室。这些资料数据量的来源分布为：历史纸面资料占50%，其他各类数据文件转换、数据迁移、外购、下载占50%。后者可以直接入库，费用和效用对于传统的和集成的全文系统是相同的，差别在于5千万字的历史纸面文档的录入费用和建设时间。在费用、人员硬约束条件下，传统全文系统需要半年以上才能完成1亿字的输入工作，而集成全文系统一般仅需2个月就可以基本完成输入工作，且传统全文系统录入5千万字历史纸面文档的方法只能采取手工录入校对，集成全文系统仅需录入少量结构化信息，其余信息只需高速扫描输入，人员数量、费用大大节省，因此肯定比传统全文系统节约录入费用，而增加的扫描设备维护费少到可以忽略不计。

信息系统发展的历史和经验表明，一个办公管理系统的寿命期一般不超过 5 年。因此，我们将系统寿命期定为 5 年。系统运行（效用）期为 5 年减信息输入期的时间。则传统全文系统的效用期间为4.5年，而集成系统的效用期为 4 年零 10 个月。信息输入期和系统运行期间传统全文系统需要二十人以上

录入校对纸面文档，而集成全文系统一般3、4人左右就可以完成扫描和结构数据录入。

传统全文系统的全寿命效用=信息输入期数据量+年增数据量×4.5年

=1亿字+1亿字/年×4.5年=5.5亿字

集成全文系统的全寿命效用=信息输入期数据量+年增数据量×4年零10个月

=1亿字+3亿字/年×4.8年=15.4亿字

表1 传统全文系统与集成全文系统全寿命费用、效用和效用期比较

	系统费用			系统效用		系统效用期（年）
	系统建设期投资费用（万元）	系统寿命期费用		信息输入期效用（亿元）	系统运行期效用（亿元）	
		系统管理维护费用（万元/5年）	系统录入校对费用（万元/5年）			
传统全文系统	20	10	250	1	4.5	4.5
集成全文系统	25	10	30	1	14.4	4.8

显然，集成全文系统不但效用期长于传统全文系统，而且全寿命效用也高于传统全文系统。在费用低于约束条件的情况下，集成系统的全寿命费效比远远低于传统系统。即便我们在系统建设期甚至以前就开始信息录入工作，但仅能提高有限的系统效用期，无法改变输入效率低、成本高的事实。

从图3可以看出，集成全文系统虽然在建设初期由于部分高速I/O、备份软硬件的投入使得费用比传统全文系统略高，但由于节约了运营费，全寿命费用上升幅度很低并在经过一个平衡点以后低于传统系统。图4反映了系统全寿命效用，由于数据的丰富使系统可用性增加，集成系统的效用在一开始就远远高于传统系统的效用，且保持较高的效用增长水平，而传统全文系统由于要有一个较长的数据输入、迁移期，在这个期间效用不明显，只有在数据量达到一定水平以后效用才开始缓慢增加，且要受数据输入费用、速度的制约，效用增长依然缓慢。

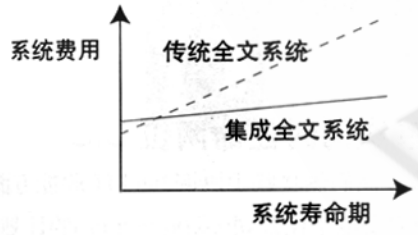

图3 系统全寿命费用比较

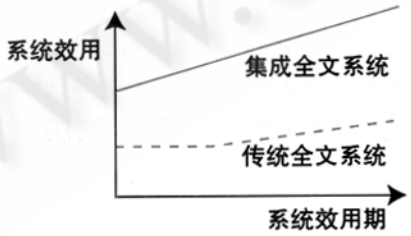

图4 系统全寿命效用比较

因此，在建设机关档案部门的全文检索系统时，必须考虑机关全文系统的应用特点，寻求最佳的系统全寿命费效比，而实现该目标的一个可取手段就是采用高速扫描、OCR和全文系统的集成应用。