基于关键字的文献索引和相似分类算法

西南民族学院　　杨宪泽

摘要：本文以文献管理系统为基础，介绍了关键字映射索引算法和相似文献分类算法。关键字映射索引算法使关键字与文献存贮地址直接映射，不实施反复比较操作，有较高效率；相似文献分类算法是智能检索的初步探讨。两算法适宜今后在计算机信息处理中广泛应用。

一、引言

以文献关键词检索文献，是计算机检索手段之一。我们在设计成果文献管理系统中，规定进入系统的文献必须具有三条以上的关键词。系统以关键词为排列的倒排文档结构实现。我们所做的工作如下：

逻辑运算。按用户的特殊检索要求。输入多个检索关键字及相联的逻辑符。逻辑运算能形成满足关键字的文献集合。

例如，检索 90 年计算机系获奖文献。这是对关键字获奖情况和单位(计算机系)检索的与运算，即 S＝获奖情况 AND 计算机系。

经检索，计算机系 1990 年有文献 41 篇，全院 1990 年获奖文献共 36 篇。其中文献 62，68，71，75 在两项检索中都出现，所以它们是 1990 年计算机系获奖文献。因此，运算完成后程序仅输出这四篇文献。

系统为逻辑运算设计了一个处理程序，算子按逆波兰式变换，其优先级是括号(1)，OR(2)，AND(3)，NOT(4)。

处理变长文献记录。存储长文献记录时，需要确定合理的记录长度。设置过长，对于短文献信息将会造成存贮空间的极大浪费；设置过短，长文献信息会被截掉一部分。为解决变长文献记录问题，可把存储的文献记录信息分成两部分：一是定长信息，供检索操作用；另一部分是变长信息，被定长信息链接，即定长信息→链指针→变长信息。

每一文献记录的定长信息指针指向自身的变长信息，若指针为空，表示这一文献记录无变长信息。

关键字映射索引。现有的检索算法可为两类：第一类完全建立在比较基础上，检索依赖于关键字个数和所进行的比较次数。这类算法二分检索效率最高，为 O(Nlog2N)；第二类称为散列算法，其核心是构造 Hash 函数使信息关键字通过计算而落在预先确定的值域，产生冲突的关键字再比较查找完成检索。实践证明，要分析关键字设计出一个好的 Hash 函数较困难，特别难以克服数据分布不均带来的障碍。如文献[1]的努力仅使数据在服从某些特殊概率的分布的条件下，检索的平均工作量为 O(N)。

我们利用计算机内每个字符可转换成 ASCII 码介于 0～255 之间的十进制数这一特点，结合文献[3，4]基本思想，提出了一个能克服数据分布不均，亦可在不同关键字情况下以相同方式建立索引的高效算法。

相似文献分类。系统按模糊检索方法[5]，设计了一个按课程实现文献相似分类的算法。

例如，n 篇文献中的一些关键字可能同属于人工智能，数据结构和数据库，为了教学和科研参考方便，我们要自动实现文献按课程分类。分类原则至少使每篇文献分到一类中去。

由于系统的大多数设计工作是现成方法的组合应用，本文不再讨论。本文要着重介绍的是关键字映射索引算法和相似文献分类算法，因为这两个算法可能对检索方法新的研究和应用有所启示。

二、关键字映射索引算法

1.索引构思

关键字 K(字符串)，可分划成单一字符 K1,K2……

Kd(d为字符个数),并能分别求出它们的ASCII码十进制值,实施命令是

LEN(k $) ：求字符串(关键字)长度

MIDS(k $,I,1) 切分字符串中第I个字符

ASC(k $) ：求一个字符(k1)的ASCII码十进制值。

实施上述命令后,对任一待检索关键字 Ki(i=1,2,…N),已分划成≤255的正整数 Ki1,Ki2,…Kid(一个字符的 ASCII 码的十进制数值≤255)。映射关系集 F{=Fi|j=d,d-1,…,}1 由下式表示:

Fd=Ki1+Ki2+…+Kid

Fd-1=Ki1+Ki2+…+Kid-1

：

F1=Ki1

映射要建立这样的关系:如果 Ki→Fd 映射是单一的,那么可建立一个文献记录(关键字为 Ki)→附加数组空间单元 R(Fd)索引关系。反之,有相同 ASCII 码十进制之和的文献关键字。这时我们试探 Ki→Fd-1 是否是单一的…,只要关键字不同,一定可以在 Ki→Fj(j≥1)找到这种单一关系。如果直到 Ki→F1 都不是单一的,R(F1)建立相同关键字索引。

2.索引算法

A1:对文献关键字 Ki(初值 i=1),切分成单一字符 Ki1,Ki2,…,Kid。

A2:求 Ki1,Ki2,…,Kid 的 ASCII 码的十进制数值,最大值为 255;并求和,产生 Fd,Fd-1,…,F1。

A3:开辟 dx255,(d-1)×255,…,1×255 个存贮空间 Rd,Rd-1,…,R1,作下列计算,

(1)让 Fd 值与 Rd 空间地址值 L 相对应,L←Fd,计数器 Pld←Pld+1(Pld,Pld-1,…,Pl1 初值赋 0)。

(2)若 Pld>1,在 Rd-1 空间中,L←Fd-1,且 PLd←Pld+1;若 Pld>1,在 Rd-2 空间中,L←Fd-2,…。如果 Pl1>1,为相同文献关键字。

(3)若 Plj=1(j≥1),把 Ki 的文献原始存贮地址填入 RLj(L)中,PL1>1 的相同关键字,采用链接方式建立索引。

A4:i←i+1,重复 A1～A3,直至 i=N 为止。

3.检索算法

索引算法中,文献关键字与 R 空间形成了唯一对应关系,为检索提供了捷径。B1:给定待检索文献关键字 K,分划成 K1,K2,…Kd。

B2:求 K 的 Fd,Fd-1,…,F1。

B3:①若Pj(Fj)>1转②;Pj(Fj)≤1转B4(j初值赋d)。

②j←j-1,j>1 转①j=1 转 B5。

B4:①Pj(Fj)=1,Rj(Fj)为索引地址检索成功。②Pj(Fj)<1,失败退出。

B5:相同关键字文献全部检索。

4.算法分析

(1)为了防止算法附加存贮空间太大,系统限制关键字字符≤4。若突破这一规定,系统分解。

关键字=关键字 1 AND 关键字 2

其中,关键字 1 字符数=4。这时,关键字 1 对应的文献,关键字 2 也对应。

(2)空间复杂性:算法附加的存贮空间在建立索引中,共 255(d+(d+1)+(d+2)+……+1)=128d(d+1)。

本系统以中文字符 d=4 计算,附加存贮开销 5100。这对于 N 很大的关键字集合来说,附加存贮空间不大。

(3)时间复杂性:最小检索长度 1,最大检索长度 d,为 4;平均检索长度据算法分析可知,在 N 个关键字均匀分布的情况下,换算到 Rd,Rd-1,…R1 空间地址值的概率相同,有 P(E)1/d。因此,平均检索长度。

$$E(N)=1\cdot 1/d+2\cdot 1/d+\cdots+d\cdot 1/d=12(i+d)/2$$

本系统平均检索长度为 2.5

(4)实施这一算法时,必须对待检索关键字 K 进行一系列变换。这是不是增加了很多操作时间?我们采用文献[6]指令流方式推算,结果这一算法对 K 的变换所需的时间,相当于折叠式 Hash 函数构造法所需的时间。这说明这一算法对 K 的变换并不比一般构造 Hash 函数所作的变换复杂。

三、相似文献分类算法

1.算法构思

设 n 篇文献由 $X_1,X_2\cdots,X_n$ 组成。n 篇文献中含有 m 个不同的关键字 $K_1,K_2,\cdots K_m$。这样,一篇文献 X_r 可用 m 维向量来描述

$$X_r(\Phi_{r1},\Phi_{r2},\cdots,\Phi_{rm})$$

其中

$$\Phi_{rj}=\begin{cases}1 & X_r \text{ 中有关键字 } K_j\\ 0 & X_r \text{ 中无关键字 } K_j\end{cases}$$

对于每一关键字 $K_1, K_2, \cdots, K_m$，事先标注它是属于某一类或多类（某一门课程或多门课程）。例如，作者发表一篇文章，给出三个关键字，只允许作者认定这篇文章属于一门课程，那么这三个关键字均属这门课程；然而，在另一篇被认定为属于另一门课程的文章中，有一个关键字与前述文章相同，那么这一关键字将属于两类。

统计 K_j 在第 i 类中出现的概率 P_{ij}

$$Pij=\frac{Lij}{Lj}\quad (j=1,\ 2,\ \cdots,\ m,\ i=1,\ 2,\ \cdots,\ d)$$

式中，L_j 是 K_j 在文献中出现的总次数，L_{ij} 是 K_j 认定属于 i 类的总次数。d 为分类数。

每一类 Gi 的模糊度量 0～1 之间的隶属函数可由下式算出

$$\mu^i(Xr)=\frac{\sum_{j=1}^{m}\Phi rjPij}{\sum_{j=1}^{m}\Phi rj}\quad (i=1,\ 2,\ \cdots,\ d\quad r=1,\ 2,\ \cdots,\ n)$$

有了隶属函数，即为相似分类提供了依据。

2.算法描述

C1：确定分类数 d。每篇文献假定属于一类（作者填表认定）。

C2：r 从 1 至 N，输入每篇文献 Xr 的关键字 K1，K2，…，Km（即建立索引）：

（1）若有 Kj 相同（j=1，2，…，m），Lj←Lj+1（Lj 初值赋 1）。

（2）若 Kj 所属文献为 i 类（i=1，2，…，d），Lij←Lij←+1（Lij 初值赋 0）。

C3：r 从 1 至 N，Xr 含有的关键字 Kj（j=1，2，3，…）对应 Φrj=1（其余 Φrj 初值已赋 0）。

C4：[初值 j=1]i 从 1 至 d，计算 Pij=Lij／Lj。

C5：j←j+1，直到 j=m，重复 C4。

C6：[初值 r=1，j=1，i=1，A=0，B=0]计算 A←A+Φrj，B←ΦrjPij。

C7：j←j+1，直至 j=m 为止，重复 C6。

C8：μir=B／A，i←i+1，直至 i=d 为止，重复 C6，C7。

C9：r←r+1，直至 r=n 为止，重复 C6～C8。

C10：[分类开始，每篇文献至少分到一类中去，初值 μ=0.9，r=1]i 从 1 至 d，若 μri≥μ，Xr→Gi(j)，j←j+1（Gi 为类集合），此时，Pr=1。

C11：r←r+1，直至 r=n 为止，重复 C10。

C12：r 从 1 至 n，若 Pr 均为 1，分类结束，否则 μ←μ−0.1，重复 C10，C11。

3.几点说明

（1）算法 C1～C9，是相似分类的一种计算机自动计算方法，关键在于确定合适的隶属函数。这里只是一种探讨，也可以采用其它方法确定隶属函数。

（2）C10～C12 实现自动分类，最后文献被分到 G1，G2，…Gd 集合中去。由于每篇文献至少分到一类中去，我们用 Pr 作为标志，Pr=1，意味着 Xr 至少进入了一类。C12 首先判断是否有 Pr=0，若有，意味着还有文献没有入类，这时我们就降低隶属函数阀值再实施分类，直到每篇文献至少进入一类为止。

（3）该算法作为初步探讨，暂时不作复杂性分析。

四、结束语

本文以检索系统设计为基础介绍了关键字映射索引算法和相似文献分类算法。笔者认为关键字映射索引算法的实际意义在于，一般情况下提高了检索效率，抛弃了一般散列检索算法为构造好的 Hash 函数必须分析不同情况关键字所面临的困难，也解决了构造 Hash 函数的散列算法难以解决的数据分布不均带来的问题。

相似文献分类算法是智能检索有意义的探索。它虽是初步的，但却有一定的研究价值，更富有实践性，易于今后被改进和应用。

参考文献：

1.AKL, S・G, and Meijer, H・, On the Average-Case Complexity of Bucketing Algorityms "Journal of Algorithms 1(1982), 9-13

*2.*周建钦等，随机分组查找算法，科学通报，第 *35* 卷，*1990* 年 *24* 期。

*3.*杨宪泽，子域映射快速排序法研究，科学通报，第 *35* 卷，*1990* 年 *15* 期。

*4.*杨宪泽，直接映射式字符检索算法，中文信息学报，第 *5* 卷，*1991* 年第 *3* 期。

*5.*冯德益等，模糊数学方法与应用，地震出版社，*1985* 年。

*6.*杨宪泽，长记录位置不变的排序算法，软件学报，*1993* 年第 *1* 期。

▲

批处理接口(OBI)进行处理：提供系统控制参数和数据分配／定义参数并进行数据完整保护处理，形成作业控制语句流(JCS)提交到 OS 的作业控制系统执行。OBI 处理需要决定本批处理应用与联机应用是否为互斥的，如数据重组／加载等。如为互斥，需要待联机应用结束后再将 JCS 提交(如图 3)。

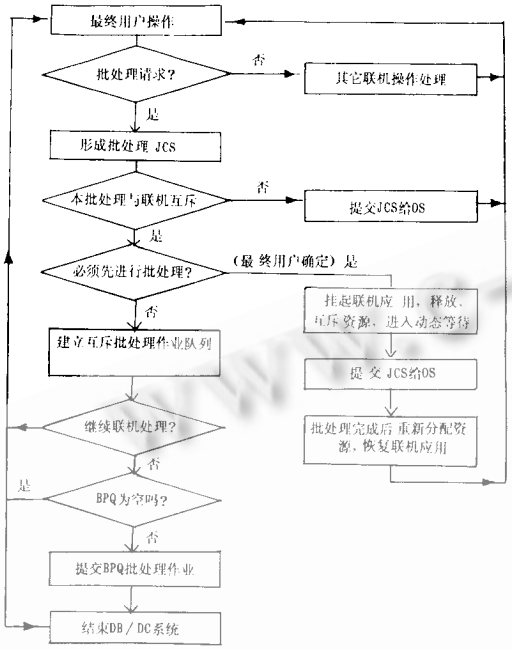

图　3

2.主机——工作站协处理

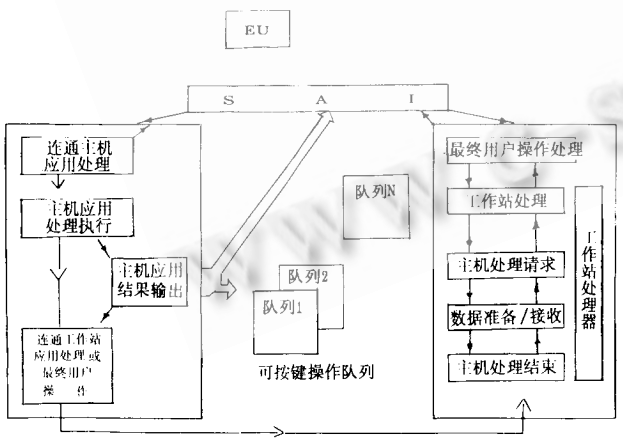

图　4

众所周知，汉字智能工作站(HIWS)在图形／汉字处理方面较大型主机要方便的多；STIMS 系统设计使 HIWS 发挥最大效能。STIMS 设计并实现了主机——工作站协处理(HOSTERM)机制：将图形分析、联机帮助及用户特殊需求编程接口支持功能分布在 HIWS 上运行，STIMS 其它功能分布在主机上运行。前者所需数据由后者运行提供，但是，在 SAI 界面上，最终用户只是体会到一个一致的界面，并无异样感觉，其结构如图 1 工作原理如图 4 所示：主机应用处理与工作站应用处理间依靠可直接访问子队列的队列设施完成可能的信息耦合，控制耦合则由应用处理程序内部逻辑确定。

三、界面设计

联机系统用户界面可认为是两类设备：输入类，即键盘(包括光笔、鼠标等)；输出类，即 CRT 显示器和打印机。那么界面设计涉及键盘用法、键盘请求处理、输出外观和人机界面对话方式等四类问题。

1.键盘用法

键盘用法主要是程序功能键(PF 键)、程序请求(PA)键、清除(CLEAR)键和其它非字母符号的用法。通过统一它们的功能用途实现键盘处理的一致性：PF1 用做 HELP 键，PF3 作用结束当前操作返回调用键；PF7、PF11、PF8 和 PF12 分别用作显示窗口屏幕动态输出和操作输入键等。在主机应用程序和工作站应用程序中，完全依 ENTER 键作处理激发器。通过控制屏幕属性限定键盘的输入区域。

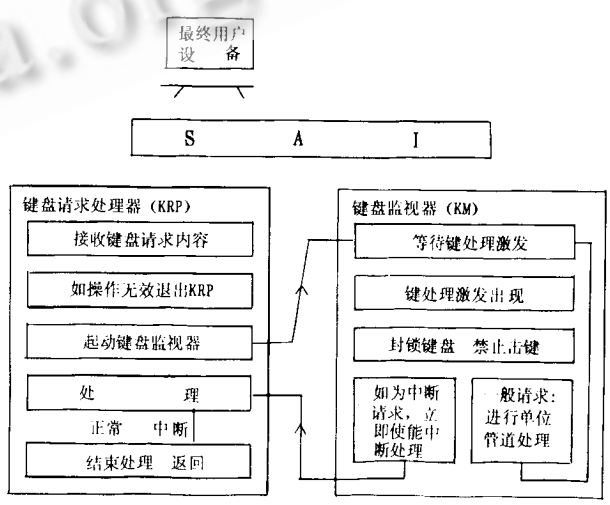

图　5

2.键盘请求处理器(KRP)

保证当前状态下激发的键盘操作请求得到适当处理,就是键盘请求处理的职责。如果 KRP 认为键盘操作请求在当前状态下是有意义的,则进行处理;如果无意义则提示最终用户操作错误并复原到原状态。当应用工作在工作站状态时,控制十分简单。应用工作在主机状态时,其处理原理如图 5 所示。

3.输出外观

屏幕由窗口使用,窗口上设置标题区、提示内容区、环境景观区和盘面体区组成。其中盘面体区是联机应用与最终用户的通讯区;环境景观区由联机应用程序提取系统服务信息,如当前日期、时间、操作路径等使用;提示内容区包括下一步可操作内容提示、功能键功能定义等,窗口要达到外观和内涵的一致性。

4.人机对话方式

使用以下人机对话方式:1 菜单选择输入;2 参数直接输入;3 修改引导诱惑参量输入;4 确认输入等。

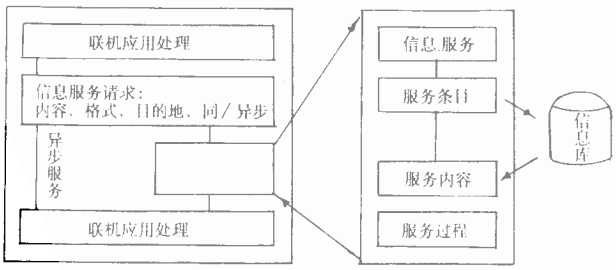

图 6

四、信息与程序无关性设计

信息(Message)是联机应用向最终用户发出的操作提示、处理结束标志、异常情况通告、误操作警告或对最终用户询问的答复。如果把它们统统写进联机应用程序中,既导致程序规模变大又不利于修改。信息与程序无关性设计即采用把信息与应用程序分开单独设计方式。STIMS 采用图 6 所示的分离信息服务方式。程序中仅置一个全系统中唯一的信息标识(MID),其体该 MID 及其所对应的内容存储在一个信息库中。当请求时,信息服务与联机应用主任务并行操作:首先根据 MID 从信息库中得到信息的固定部分,补充动态部分,发向原请求程序提供的目标地。

五、动态程序设计

大型联机信息管理系统无疑是多用户系统,其应用处理程序支持多道程序运行是基本要求。STIMS 采用了"可重用与可刷新"和缓冲动态分配技术来支持。

1.可重用与可刷新编程

为了在有限的运行空间内提供尽可能多的并行操作,STIMS 的程序模块在运行空间仅保存一份映象。每个联机应用都使用这同一份程序运行。所以程序设计时严格区分程序执行部分和数据区、变量区。后者在程序执行时动态获得:需要时由应用去请求 OS 动态分配对应本程序当前运行的数据/缓冲区,放弃时通知 OS 撤消(图 7)。这样首先保证了可刷新特性。DB/DC CICS/VS 设施保护多道程序设计的断点从而保证了其可重用性。

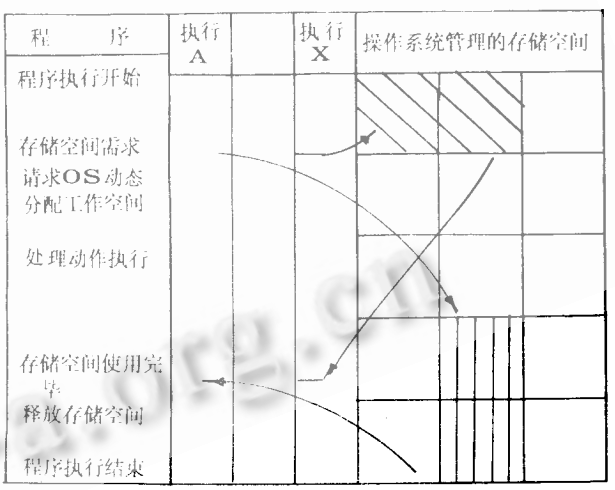

图 7

2.缓冲动态分配

在联机处理中需要大数据空间(如设置窗口缓冲)时往往不宜占用供程序运行使用的存取空间。此时应使用外存储设备空间(如 DASD)来提供。STIMS 在 DASD 中开辟一个池(POOL)作为可按键操作的队列。依据不同最终用户的标识(ID)和当前应用处理的子功能号组合形成使用队列的键保证不同最终用户间或/和最终用户不同子操作间互不干扰,并行运行。