[19] 中华人民共和国国家知识产权局

[51] Int. Cl[7]
G06F 17/30

[12] 发明专利申请公开说明书

[21] 申请号 01821280.8

[43] 公开日 2004 年 3 月 17 日

[11] 公开号 CN 1483169A

[22] 申请日 2001.12.19 [21] 申请号 01821280.8
[30] 优先权
[32] 2000.12.29 [33] US [31] 60/258,991
[86] 国际申请 PCT/US01/49260 2001.12.19
[87] 国际公布 WO02/054289 英 2002.7.11
[85] 进入国家阶段日期 2003.6.23
[71] 申请人 国际商业机器公司
地址 美国纽约
[72] 发明人 D·卡梅尔 D·科亨 R·费金
E·法尔基 M.赫尔什科维奇
Y·马雷克 A·索弗

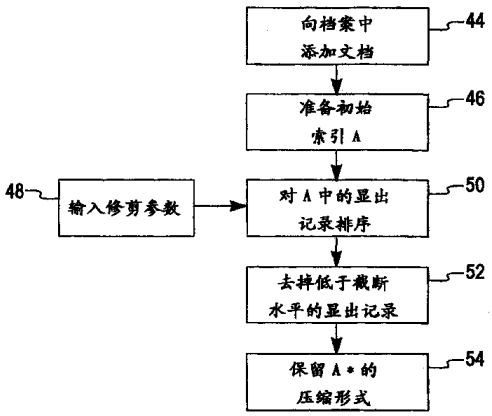
[74] 专利代理机构 北京市中咨律师事务所
代理人 于 静 李 峥

权利要求书 2 页 说明书 9 页 附图 2 页

[54] 发明名称 有损索引压缩

[57] 摘要

本发明提供了一种用于实现一种方法(图 2)以修剪一个文本文档语料库的索引的装置，其中该方法包括以下步骤：对索引中的显出记录进行排序(50)和从该索引中修剪掉(48)排序中低于给定水平的显出记录。本发明的这些修剪方法是有损的，因为一些文档显出记录被从完全的索引中去掉；然而，用户无法区别该有损索引和完全索引。



ISSN1008-4274

1. 一种用于为文本文档语料库建立索引的装置，其特点在于一个索引处理器，它被安排成建立文档中出现的项的倒排索引，该索引包含文档中那些项的显出记录，该处理器进一步被安排成建立该索引中显出记录的排序，并从该索引中修剪掉排序中低于给定水平的显出记录。

2. 根据权利要求 1 的装置，其中处理器被安排成对至少是一些项的每一项分别确定单独的排序，并对这至少是一些项的每一项修剪其单独的排序。

3. 根据权利要求 2 的装置，其特点还在于一个用户界面用于接收至少一个参数，其中该处理器被安排成根据该参数和该单独索引排序设置给定水平。

4. 根据权利要求 3 的装置，其中该至少一个参数包含要从索引中检索出的文档数 k 和在一个查询中允许的项数 γ，而且其中该处理器被安排成根据从排序顶点算起排序为 k 的一个文档的评分来设置该水平。

5. 根据权利要求 4 的装置，其中该处理器被安排成通过以 γ 除这一文档的评分来设置给定水平。

6. 根据权利要求 3 的装置，其中该至少一个参数包含从该排序中检索的部分文档的个数 δ 以及在一个查询中允许的项数 γ，而且其中该处理器被安排成根据该排序中的那些文档中的第一个文档的评分、δ 以及 γ 来设置给定水平。

7. 根据权利要求 6 的装置，其中处理器被安排成将此第一个文档的评分乘以 δ 和除以 γ。

8. 根据权利要求 1 的装置，其中处理器被安排成根据搜索空间中查询对于文档显出记录的统计分布信息来选择要修剪的显出记录。

9. 根据权利要求 8 的装置，其特点还在于一个用户界面用于接收至少一个参数，其中该处理器被安排成根据该参数和索引排序来设置给定水平。

10. 根据权利要求 9 的装置，其中该至少一个参数包含：在修剪过的倒排索引中要保留的评分个数 M。

11. 根据权利要求 10 的装置，其中处理器被安排成确定至少是一些项的概率并将这至少是一些项的每一项的显出评分乘以该项的概率，并且以乘过的显出评分对所有显出记录排序，其中给定水平包含从排序顶端算起的文档 M 的评分。

12. 根据权利要求 1 的装置，其中该索引处理器的特点在于一个具有大存储容量的计算机并包含用于将修剪后的索引传送到具有有限存储容量的设备的装置。

13. 根据权利要求 12 的装置，其中具有有限存储容量的设备包括手持计算设备。

14. 一种用于实现对文本文档语料库建立索引的方法，其中该方法的特点在于如下步骤：

建立文档中出现的项的倒排索引，该索引包含这些项在文档中的显出记录；

对索引中的显出记录排序；以及

从索引中修剪掉排序中低于给定水平的显出记录。

有损索引压缩

技术领域

本发明一般地涉及在大的文本数据体中进行计算机化搜索的方法和系统，具体地涉及建立搜索索引。

背景技术

快速和精确的文本搜索引擎被广泛地用于网络和桌面应用。正在出现的手持设备，如 Palm Pilot™，具有足够的存储能力，允许在设备上存储完整的中等大小的文档集以供快速引用和浏览。希望以高级的基于索引的搜索引擎装备这些设备，但手持设备上的存储能力仍然相当有限。

大多数高级信息检索（IR）应用创建倒排索引，以支持对给定文档集的高质量搜索服务。这类系统的一个实例是 Guru 搜索引擎，它由 Maarek 和 Smadja 在“基于词法关系的全文本索引，一个应用：软件库”（关于信息检索的研究与开发的第 12 届国际 ACM-SIGIR 年会文集，第 198-206 页，1989）一文中予以描述，这里以引用方式将全文纳入。对文档集内的每个文档进行分析并根据该文档的内容由索引单元或索引项的向量简表来表示。一个索引项可以是一个词（word）、一对紧密相关的词（词法结合体）或一个短语。在一个文档中的每个索引项与它所关联的显出表（posting list）一起存储在索引中。

显出表包含显出记录，这里每个显出记录包括含有该索引项的文档的标识符，该索引项在那个文档中的评分，还可能有关于该索引项在该文档中出现情况的附加信息，如出现次数和出现位置的偏移量。在许多信息检索系统中使用的一个典型评分模型是 tf-idf 公式，由 Salton 和 McGill 在“现代信息检索引论”（McGraw-Hill 出版社，1983）中描述，该文在这

里以引用方式将全文纳入。项 t 对文档 d 的评分依赖于 t 在 d 中的项频度（tf）、文档 d 的长度以及在该集合中含 t 的文档个的倒数（idf）。

Chris Buckley 等在“使用 SMART 的新检索途径: TREC 4”一文（第四届文本检索会议（TREC 4）文集，第 25-48 页，Gaithersberg, Maryland, 1995 年 11 月）中描述了一个 tf-idf 公式示例，该文在这里以引用方式将全文纳入。该公式给出，文档 d 对项 t 的评分 A(t, d)是

$$A(t,d)=\frac{\log(1+tf)}{\log(1+avg_{tf})}\times\log(N/Nt)/|d|$$

这里 avg_{tf} 是文档 d 中的平均项频度，N 是在该集合中的文档数，Nt 是含有项 t 的文档数，|d|是文档 d 的长度。通常，|d|是由 d 中(唯一) 项的数量的平方根来近似的。

在搜索时，从用户查询中提取出项，从倒排索引中检索出这些项各自的显出表。通过对从属于同一文档的显出记录的评分求和，积累该文档的显出评分以形成文档评分。在这一过程结束时，这些文档按它们的评分排队，并返回具有顶级评分的那些文档。

对大的文档集合建立索引造成难于维护的巨大的索引文件。在索引压缩领域已做了大量工作，以便得到较小的索引文件。在本领域存在两种互补的途径。一种途径是在数据结构层进行压缩，即保留所有索引数据而同时试图得到显出表的更紧凑表示。另一种途径是通过删除或组合项，例如省去无用词（stop-word），以及潜在语义索引（LSI）来修剪索引。这类索引修剪的主要目的是通过从索引项中去掉可能降低搜索精度的那些项来降低索引系统中的“噪声”，但它对减小索引尺寸的实际作用使它与索引压缩这一主题密切相关。

在省去无用词时，使用语言统计找出在语言中出现如此频繁以致在大多数文档中不可避免地会出现的那些词。在构成倒排索引时，在该语言中很频繁出现的那些词（无用词）被忽略。诸如“the”和“is”等词对检索任务没有贡献。如在“第七届文本检索会议（TREC-7）概述”（第七届文

本检索会议（TREC-7）文集，国家标准和技术研究所，1999）中呈现的那样，TREC 集合列举了在一般性文本文档中的词频度。该文在此以引入方式纳入。通过忽略 TREC 集合中的 135 个最频繁出现的词，发现有大约 25％显出记录被去掉（Witten 等，“管理数千兆字节”，Morgan Kaufman Publishers, San Francisco, California, 1999,该文在此以引用方式纳入）。

潜在语义索引（LSI）由例如 Deerweester 等在“利用潜在语义分析建立索引”（美国信息科学杂志，第 41 卷第 1 期（1990）第 391-407 页）一文中做了描述，该文在这里以引用方式纳入。LSI 使用称作“奇异值分解”（SVD）的统计技术把倒排索引表示成三个矩阵的乘积。这一表达式通过保留最有意义的那些项去掉所有其他项来减少索引中的项数。LSI 和省去无用词都是以项为粒度进行操作。换言之，它们只能从索引中修剪掉整个项，于是，如果某项一旦被修剪掉，该项便根本不出现在索引中。当一项被修剪时，它的整个显出表被从索引中去掉。

动态修剪技术在索引已被建成之后在文档排序过程中确定某些项或文档显出记录是否值得加入到累积文档评分中以及该排序过程应该继续还是停止。Persin 在“用于快速排序的文档过滤”（关于信息检索的研究与开发的第 17 届国际 ACM-SIGIR 年会文集，Dublin, Ireland, 1994, SIGIR 论坛专集，第 339-348 页）一文中描述了这类技术的示例，该文在这里以引用方式纳入。动态技术应用于给定的查询，从而减少查询时间。动态技术对索引的大小没有影响，因为它们应用于已经存储的索引。

发明内容

在本发明的优选实施例中，把一个集合中的项与文档关联起来的倒排索引是在文档显出记录粒度级进行修剪的，而不是像在本领域已知的系统中那样在项级粒度上进行修剪。如下文中描述的那样，通过对给定项适当选择要修剪的显出记录，索引的大小能被显著地减少而从用户的观点看又不会显著地影响索引的搜索精度。

优选地，为文档的显出记录确定矩阵，然后将矩阵用于选择要从倒排索引中去掉的显出记录。应用这些矩阵的方式是要使得当用户以给定的查询来搜索被压缩的倒排索引时，返回的文档列表与在未被修剪的索引中由同样查询返回的顶级文档列表基本相同。本发明的修剪方法是有损的，因为某些文档显出记录被从索引中去掉了，这与本领域已知的方法不同，那些方法通过使用紧凑的数据结构和表示把数据存储在显出表中来压缩索引。有损和无损方法能彼此互补。在以有损方式修剪索引后，该索引能进一步以无损方式压缩，从而得到比单独使用这两种方法中任何一种可能得到的还要小的索引。

所以，根据本发明的一个优选实施例，提供了一种装置用于实现对文本文档语料库建立索引的方法，包括如下步骤：

建立文档中出现的项的倒排索引，该索引包括这些项在文档中的显出记录；

对索引中的显出记录排序；以及

从索引中修剪掉排序中低于给定水平的显出记录。

对显出记录排序可以包括对至少是一些项的每一项分别确定单独的排序，而对索引的修剪可以包括对这至少是一些项的每一项修剪其单独的排序。

优选地，修剪该索引包括从用户接收至少一个参数并根据该参数和单独的索引排序来设置给定的水平。

再有，这至少一个参数优选地包括要从索引中检索出的文档数 k 和在一个查询中允许的项数 γ，而设置给定水平包括根据从排序顶点算起排序为 k 的一个文档的评分来设置该水平。

根据一个实施例，设置给定水平优选地包括以 γ 除这一个文档的评分。

在另一个实施例中，这至少一个参数包括从该排序中检索出的部分文档的个数 δ 以及在一个查询中允许的项数 γ，而设置给定水平包括根据该排序中的那些文档中的第一个文档的评分、δ 以及 γ 来设置该水平。

优选地，设置给定水平包括将此第一个文档的评分乘以 δ 和除以 γ。

在另一个实施例中，修剪索引包括根据搜索空间中查询对于文档显示记录的统计分布信息来选择要修剪的显出记录。

修剪索引可以包括从用户接收至少一个参数和根据该参数以及索引排序来设置给定水平。

这至少一个参数可以包括：在修剪过的倒排索引中要保留的评分个数M。

优选地，选择显出记录包括确定至少是一些项的概率并将这至少是一些项的每一项的显出评分乘以该项的概率，而对索引排序包括以乘过的显出评分对所有显出记录排序，而给定的水平包括从排序顶端算起的文档 M 的评分。

在一个优选实施例中，建立索引包括在具有大存储容量的计算机上建立索引并将修剪后的索引传送到具有有限存储容量的设备。

优选地，该有限存储容量的设备包括手持计算设备。

附图说明

由下文中结合附图对其优选实施例的详细描述，将会更充分地理解本发明。这些附图是：

图 1 是根据本发明的优选实施例建立搜索索引的系统的示意性图示说明；

图 2 是示意性说明根据本发明优选实施例的压缩索引方法的流程图；以及

图 3 是示意性显示根据本发明优选实施例的方法（图 2）中所用输入修剪参数技术的详细情况。

具体实施方式

图 1 是根据本发明的优选实施例建立压缩的搜索索引的系统的示意性说明。用户 10 使用一个索引处理设备 12 访问一文档档案 14，从文档档案 14 中检索出的文档可以与设备 12 上现存的文档档案组合在一起。设备 12

使用下文中详细描述的方法建立压缩的倒排索引 22。通常，被压缩的索引或档案 22 被传送到计算设备 24。设备 24 与设备 12 的区别在于它存储大索引的能力有限，优选地，用于建立索引的文档档案也被传送到设备 24。于是用户能使用设备 24 构成进入该文档档案的查询并检索出适当文档的列表，尽管装置 24 的存储能力有限。

通常，设备 12 包含一台桌面计算机或服务器，而设备 24 是一台便携式普及运算设备，如掌上设备或手持计算机，如图中所示。然而，设备 24 也可以包含桌面计算机或其他计算机工作站。

图 2 是流程图，示意性说明根据本发明优选实施例的建立压缩索引 22 的方法。这一方法的步骤优选地由设备 12 上运行的适当软件来实现。该软件可以以电子形式提供给设备 12，通过网络下载，或者在有形介质，如 CD-ROM 或非易失存储器上提供。

在文档添加步骤 44，用户 10 建立文档档案 14 或向已存在的档案添加文档。在索引准备步骤 46，如本领域已知的那样，索引压缩软件通过从每个文档中提取项，为每个文档中的每一项建立文档显出记录以及在索引中列出文档显出记录来建立初始索引 A。

每个文档显出记录带有一个评分，如在背景技术部分中描述的那样。在本领域已知计算评分的各种方式，而选择哪种方式对本发明不是至关重要的。相反，采用“如果 t 不在 d 中则 A(t,d)=0，否则 A(t,d)> 0”就足够了。

然后，在参数输入步骤 48，用户输入修剪参数。这些参数用于在索引排序步骤 50 中对索引 A 中的显出记录排序。

确定显示记录排序中的截断水平，它满足修剪参数的条件。对于给定项，所有在排序中低于截断水平的显出记录被从索引 A 中删除。在显出记录去掉步骤 52 中，以这种方式建立压缩后的索引，称作索引 $A^{'}$。利用本领域已知的项修剪和数据结构压缩方法，如在背景技术中描述的那些方法，这一索引可被进一步缩小。在索引存储步骤 54，索引的压缩版本 $A^{'}$被作为压缩索引 22 存储。

从用户的观点看，压缩索引 A^* 与原始索引 A 是完全相同的。当用户查询索引 A 或 A^* 时，他收到一个文档列表，这些文档按照它们与查询项的关系排序，这种关系是以这些项的显出表确定的。通过在步骤 48 适当地选择修剪参数和在步骤 50 和 52 应用这些参数，可以保证响应该查询由 A^* 返回的文档列表以及该列表中的文档顺序将与由 A 返回的列表的顶部基本相同。这通常是列表中用户感兴趣的那部分。在这个意义上，本发明的方法类似于图像和声音的有损压缩方法，这里数据量的显著减少是通过牺牲细节来实现的，这些细节大部分是用户察觉不到的。

现在将描述指定输入参数(步骤 48)和应用这些参数（步骤 50 和 52）的三种优选的方法。前两种方法删除尽可能多的文档显出记录，而保持为响应查询由修剪后索引返回的顶级回答尽可能地接近于由原始索引返回的顶级回答。该接近度由使用顶级回答矩阵来测量，而顶级回答矩阵是由原始索引返回的顶级结果组与修剪后的索引返回的顶级结果组之间的相似性确定的。

第三种方法是均一文档显出记录修剪法，它去掉为达到给定索引大小必须去掉的那么多文档显出记录，而又保持预期误差尽可能小。预期误差是用一个矩阵测量的，该矩阵定义为由原始的和压缩后的索引为每个查询返回的文档评分之差对所有可能的查询求和。

如果对任何给定的查询，由压缩索引和原始索引返回的对该查询的“顶级回答”完全相同，则该压缩索引被定义为与原始索引完全相同。本发明的两个优选实施例从两种可能的测量导出“顶级回答”。

· “k 顶级回答”法把“顶级回答”定义为对一个查询有最高评分的 k 个文档，这里 k 是在步骤 48 输入的。定义 γ 为任何查询中允许的最多项数。对每一项 t，值 $A(t,d_0)$，$A(t,d_1)$，...根据它们的大小在步骤 50 排序。设 Zt 为该排序中第 k 项的大小。于是在步骤 52，如果 $A(t,d)<Zt/\gamma$，则 $A^*(t,d)$ 设为 0，否则 $A^*(t,d)=A(t,d)$。$A^*(t,d)=0$ 的显出记录当然从索引中被去掉。

· “δ 顶级回答”法利用对一给定查询在从评分函数的顶级评分算起

的距离上的一个阈值来定义“顶级回答”，这里δ是在步骤48的输入。例如，如果δ=0.9，则其评分高于顶级评分90%的任何文档被认为是一个顶级回答。这里也是在步骤50对A(t,d)排序。在步骤52，对每一项t，找出最大值max(A(t,d))。设Zt=δ×max(A(t,d))。于是，如果A(t,d)<Zt/γ，则A'(t,d)=0，否则A'(t,d)=A(t,d)。A'(t,d)=0的显出记录当然从索引中被去掉。

图3扩展修剪参数输入步骤48，以用于上文提到的第三种方法，即均一显出记录修剪法。在修剪参数输入步骤55，由一个外部过程确定全部可能查询集合的概率分布$Dist_q$作为输入到系统中的一个分布。$Dist_q$可以从例如该语言中各项的分布、从一个搜索引擎的查询日志文件或从任何其他适当的方法中得到。在索引中项的分布$Dist_t$是在确定步骤56从查询和$Dist_q$导出的。项分布反映一个项t将在提交给搜索引擎的一个查询中出现的概率。一项出现的概率可用查询概率表示为$Pr(t)=\Sigma_{q\in Q, t\in q}Pr(q)$，这里Q是全部可能查询集合。在输入步骤58，用户输入希望在索引A'中保留的显出记录的个数M。然后，索引压缩的第三优选实施例在步骤50对A的值排序，并在步骤52按如下步骤建立A'：首先，根据A和$Dist_t$建立评分索引A，A(t,d)=Pr(t)A(t,d)。在A中对所有评分排序，并且确定z以便在A中刚好有M个评分大于z。请注意，在这一方法中，z是在A上的全局参数，而不是如上述前两个方法中那样对每项t有一个z。于是，如果A(t,d)<z，则A'(t,d)=0，否则A'(t,d)=A(t,d)。

本发明的发明者们已使用TREC中给出的洛杉矶时报(Los-Angles Times)数据作为经验数据对这三种方法进行了测试，这组数据包含约132,000个文档。为了改善方法性能，原始索引被修改。对每一项，在对那项的所有文档显出记录中的最小评分被从所有其他评分中减掉。对上述方法进行这一校正之后，顶级k修剪法允许修剪掉多达25%的文档显出记录，在对每项使用顶级10个评分和不超过10项的查询时不会明显降低搜索结果的质量。顶级δ修剪法允许修剪掉多达20%的文档显出记录，在对每项使用顶级70%评分和不超过10项的查询时不会明显降低搜索结果的

质量。对于所选择的文档档案，顶级 K 和顶级 δ 两种方法的表现都优于均一显出记录修剪法。

产业上的可应用性

本发明能通过提供例如用于索引文本文档语料库的装置使其在产业界得到利用，该装置包括一个索引处理器，它被安排成建立文档中出现的项的倒排索引，该索引包括文档中那些项的显出记录，该处理器进一步被安装成建立该索引中显出记录的排序，并从该索引中修剪掉排序中低于给定水平的显出记录。

还可以根据本发明的优选实施例通过提供计算机软件产品来实现和使用本发明，该计算机软件产品用于索引文本文档语料库，该软件产品包括一个计算机可读介质，其中存储程序指令，当由计算机读取这些指令时，这些指令使计算机建立文档中出现的项的倒排索引，该索引包括文档中那些项的显出记录，这些指令进一步使计算机对索引中的显出记录排序并从索引中修剪掉排序中低于给定水平的显出记录。

应该理解，上文描述的实施例是以举例方式列举的，而且本发明不限于上文中已具体显示和描述的内容。相反，本发明的范围包括上文描述的各种特征的组合和次级组合及其各种改变和修改，对于阅读过上文的本领域技术人员而言，这些都是会发生的，而且没有在现有技术中公开说明过。

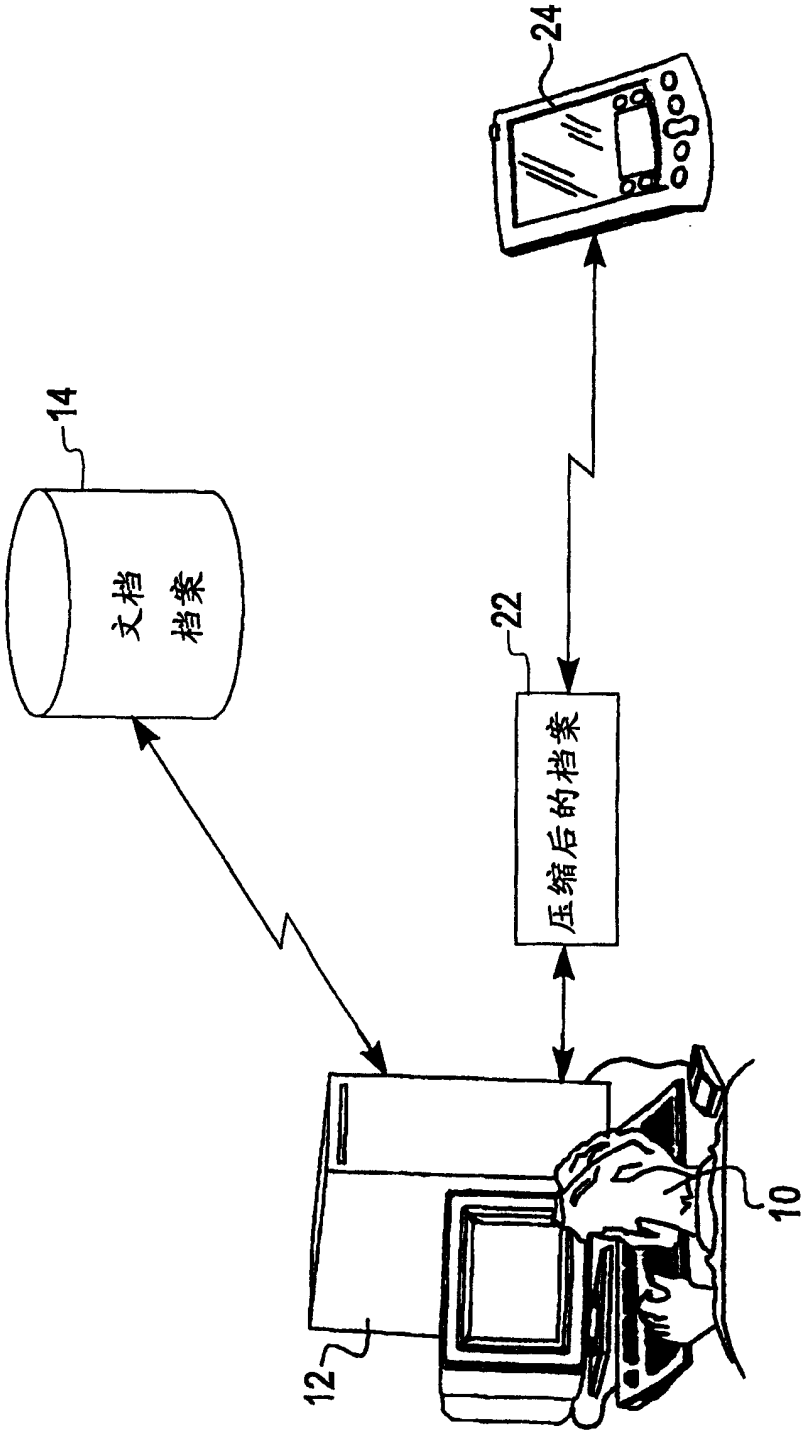


图1

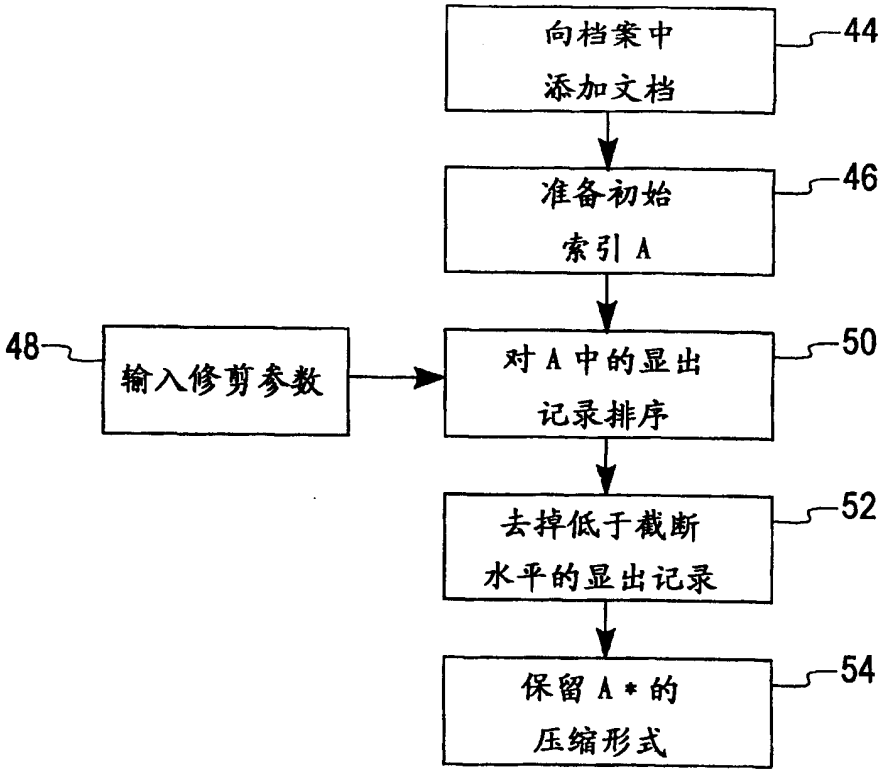

向档案中
添加文档
44
准备初始
索引 A
46
48
输入修剪参数
对 A 中的显出
记录排序
50
去掉低于截断
水平的显出记录
52
保留 A * 的
压缩形式
54

图2

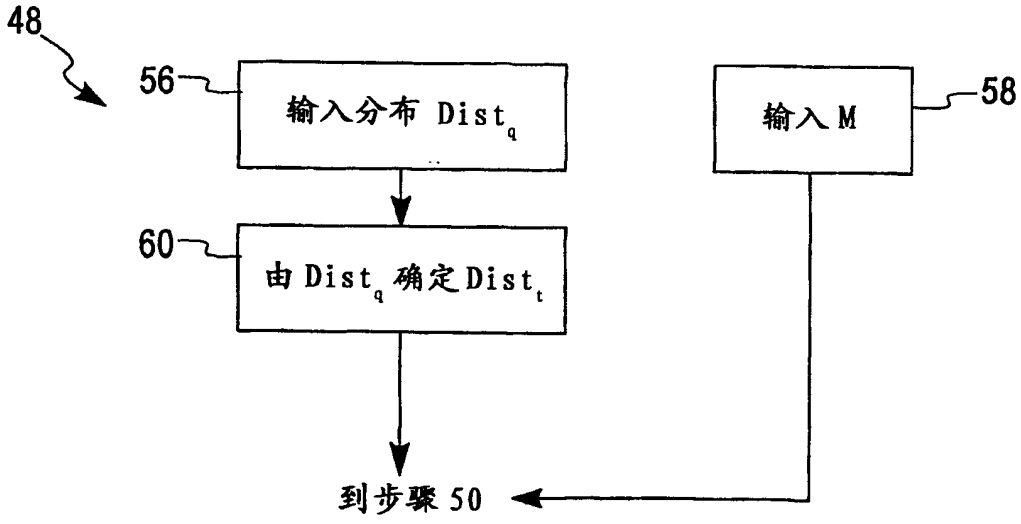

48
56
输入分布 $Dist_q$
输入 M
58
60
由 $Dist_q$ 确定 $Dist_t$
到步骤 50

图3