(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 103309951 B
(45)授权公告日 2016.08.10

(21)申请号 201310193569.X
(22)申请日 2013.05.23
(73)专利权人 北京大学
地址 100871 北京市海淀区颐和园路5号
(72)发明人 段凌宇 王哲 林杰 杨爽 黄铁军 高文
(74)专利代理机构 北京同立钧成知识产权代理有限公司 11205
代理人 刘芳
(51)Int.Cl.
G06F 17/30(2006.01)
(56)对比文件
US 2006/0004728 A1,2006.01.05,
CN 102117337 A,2011.07.06,

审查员 李萌

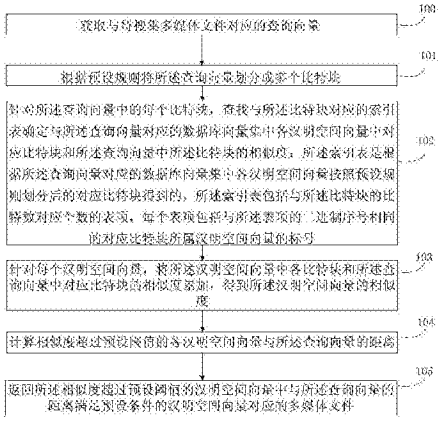
权利要求书2页 说明书8页 附图3页

(54)发明名称

在网上搜索多媒体文件的方法和装置

(57)摘要

本发明提供一种在网上搜索多媒体文件的方法和装置，将查询向量分成多个比特块后，根据对应的各比特块之间的相似度，确定数据库向量集中汉明空间向量的相似度，从而仅计算相似度超过预设阈值的各汉明空间向量与查询向量的距离并且返回所述相似度超过预设阈值的汉明空间向量中与所述查询向量的距离满足预设条件的汉明空间向量对应的多媒体文件使绝大多数检索的目标向量被包含在所述相似度超过预设阈值的汉明空间向量中，保证了检索的正确率；而且无需在整个数据库向量中对所有汉明空间向量进行遍历计算，降低了计算的复杂度，减轻了计算对系统资源的占用，可在短时间内在大规模数据库中检索出用户所需的多媒体文件，提高了检索效率。



CN 103309951 B

1.一种在网上搜索多媒体文件的方法，其特征在于，包括：

获取与待搜集多媒体文件对应的查询向量；

根据预设规则将所述查询向量划分成多个比特块；

针对所述查询向量中的每个比特块，查找与所述比特块对应的索引表确定与所述查询向量对应的数据库向量集中各汉明空间向量中对应比特块和所述查询向量中所述比特块的相似度；所述索引表是根据所述查询向量对应的数据库向量集中各汉明空间向量按照预设规则划分后的对应比特块得到的，所述索引表包括与所述比特块的比特数对应个数的表项，每个表项包括与所述表项的二进制序号相同的对应比特块所属汉明空间向量的标号；

针对每个汉明空间向量，将所述汉明空间向量中各比特块和所述查询向量中对应比特块的相似度累加，得到所述汉明空间向量的相似度；

计算相似度超过预设阈值的各汉明空间向量与所述查询向量的距离；

返回所述相似度超过预设阈值的汉明空间向量中与所述查询向量的距离满足预设条件的汉明空间向量对应的多媒体文件；

其中，所述针对所述查询向量中的每个比特块，查找与所述比特块对应的索引表确定各汉明空间向量中对应比特块和所述查询向量中所述比特块的相似度，包括：

分别确定与所述比特块的距离为r的表项序号，r为大于等于0不大于di的整数；

根据所述距离r，以及所述距离r对应的相似性因子，得到所述表项序号指向的表项对应的汉明空间向量中对应比特块和所述查询向量中所述比特块的相似度，所述距离r对应的相似性因子与所述距离r成反比。

2.根据权利要求1所述的方法，其特征在于，所述获取与待搜集多媒体文件对应的查询向量之前还包括：

根据预设规则对所述数据库向量集中各汉明空间向量进行划分，得到各汉明空间向量的比特块；

根据在各自所属各汉明空间向量中划分得到的对应位置相同顺序的各比特块，建立对应的索引表，所述对应的索引表包括2^{di}个表项，其中di为对应的每个所述比特块的比特数，二进制序号为j的表项包括所述各比特块中与j相同的对应比特块所属汉明空间向量的标号。

3.根据权利要求2所述的方法，其特征在于，所述r依次取0至x之间的每个值，x为一小于di的预设值；

所述针对每个汉明空间向量，将所述汉明空间向量中各比特块的相似度累加，得到所述汉明空间向量的相似度之前，还包括：

若一个汉明空间向量中，存在至少一个比特块未得到和所述查询向量中对应比特块的相似度，则将所述一个汉明空间向量中所述比特块和所述查询向量中对应比特的相似度设置为0。

4.根据权利要求1-3中任一所述的方法，其特征在于，所述相似度超过预设阈值的汉明空间向量中与所述查询向量的距离满足预设条件的汉明空间向量有至少两个；所述返回所述相似度超过预设阈值的汉明空间向量中与所述查询向量的距离满足预设条件的汉明空间向量对应的多媒体文件，包括：

根据所述满足预设条件的汉明空间向量对应的多媒体文件形成查询结果列表，所述查

询结果列表中多媒体文件按照对应汉明空间向量与所述查询向量的距离从小到大的顺序排列；

返回所述查询结果列表。

5.一种在网上搜索多媒体文件的装置，其特征在于，包括：

获取模块，用于获取与待搜集多媒体文件对应的查询向量；

第一划分模块，用于根据预设规则将所述查询向量划分成多个比特块；

第一计算模块，针对所述查询向量中的每个比特块，查找与所述比特块对应的索引表确定与所述查询向量对应的数据库向量集中各汉明空间向量中对应比特块和所述查询向量中所述比特块的相似度；所述索引表是根据所述查询向量对应的数据库向量集中各汉明空间向量按照预设规则划分后的对应比特块得到的，所述索引表包括与所述比特块的比特数对应个数的表项，每个表项包括与所述表项的二进制序号相同的对应比特块所属汉明空间向量的标号；

第二计算模块，用于针对每个汉明空间向量，将所述汉明空间向量中各比特块的相似度累加，得到所述汉明空间向量的相似度；

第三计算模块，用于计算相似度超过预设阈值的各汉明空间向量与所述查询向量的距离；

返回模块，用于返回所述相似度超过预设阈值的汉明空间向量中与所述查询向量的距离满足预设条件的汉明空间向量对应的多媒体文件；

其中，所述第一计算模块包括：

确定单元，用于分别确定与所述比特块的距离为r的表项序号，r为大于等于0不大于di的整数；

评分单元，用于根据所述距离r，以及所述距离r对应的相似性因子，得到所述表项序号指向的表项对应的汉明空间向量中对应比特块和所述查询向量中所述比特块的相似度，所述距离r对应的相似性因子与所述距离r成反比。

6.根据权利要求5所述的装置，其特征在于，所述装置还包括：

第二划分模块，用于根据预设规则对所述数据库向量集中各汉明空间向量进行划分；

索引表建立模块，用于根据各汉明空间向量划分得到的对应顺序的各比特块，建立对应的索引表，所述对应的索引表包括2^{di}个表项，其中di为对应的每个所述比特块的比特数，二进制序号为j的表项包括所述各比特块中与j相同的对应比特块所属汉明空间向量的标号。

7.根据权利要求6所述的装置，其特征在于，所述r依次取0至x之间的每个值，x为一小于di的预设值；还包括：

设置单元，用于若一个汉明空间向量中，存在至少一个比特块未得到和所述查询向量中对应比特块的相似度，则将所述一个汉明空间向量中所述比特块和所述查询向量中对应比特的相似度设置为0。

8.根据权利要求5-7中任一所述的装置，其特征在于，所述返回模块，具体用于：

根据所述满足预设条件的汉明空间向量对应的多媒体文件形成查询列表，所述查询结果列表中多媒体文件按照对应汉明空间向量与所述查询向量的距离从小到大的顺序排列；

返回所述查询结果列表。

在网上搜索多媒体文件的方法和装置

技术领域

[0001] 本发明实施例涉及计算机领域，尤其涉及一种在网上搜索多媒体文件的方法和装置。

背景技术

[0002] 现有技术中，人们通常在网上搜索多媒体文件，而搜索引擎通过输入的关键词在相应的数据库中查找，根据查找的结果向人们推荐与要搜索的多媒体文件表达意义相同和相近的多媒体文件，现有的多媒体文件的特征通常由汉明(hamming)空间向量表示，而通过汉明空间向量间的距离就可判定多个多媒体文件物体间的相似度。

[0003] 以图像搜索为例，两幅图像的相似度取决于对应的汉明空间向量间的汉明距离；对于一幅查询图像，首先提取表达该查询图像视觉特性的汉明空间向量作为查询向量，随后计算该查询向量与数据库中所有图像对应的汉明空间向量的汉明距离，最后把距离该查询向量最近的若干个向量对应的数据库中的图像作为搜索结果返回。这个问题的本质就是汉明空间向量的K近邻查询问题。现有技术通常采用遍历的方式计算查询向量与数据库中每个汉明空间向量之间的汉明距离，然后从中找出距离查询向量最近的若干个汉明空间向量，将这些汉明空间向量对应的多媒体文件返回给用户。

[0004] 上述现有技术的不足之处在于：现有的遍历方式的时间会随着数据规模线性增长，随之计算量增大，导致系统资源被过多的占用；而且会导致搜索时间冗长，无法在短时间内在大规模数据库中检索出用户所需的多媒体文件，具有检索效率低的问题。

发明内容

[0005] 为克服上述缺陷，本发明实施例提供一种在网上搜索多媒体文件的方法和装置。

[0006] 第一方面，本发明实施例提供一种在网上搜索多媒体文件的方法，包括：

[0007] 获取与待搜集多媒体文件对应的查询向量；

[0008] 根据预设规则将所述查询向量划分成多个比特块；

[0009] 针对所述查询向量中的每个比特块，查找与所述比特块对应的索引表确定与所述查询向量对应的数据库向量集中各汉明空间向量中对应比特块和所述查询向量中所述比特块的相似度；所述索引表是根据所述查询向量对应的数据库向量集中各汉明空间向量按照预设规则划分后的对应比特块得到的，所述索引表包括与所述比特块的比特数对应个数的表项，每个表项包括与所述表项的二进制序号相同的对应比特块所属汉明空间向量的标号；

[0010] 针对每个汉明空间向量，将所述汉明空间向量中各比特块和所述查询向量中对应比特块的相似度累加，得到所述汉明空间向量的相似度；

[0011] 计算相似度超过预设阈值的各汉明空间向量与所述查询向量的距离；

[0012] 返回所述相似度超过预设阈值的汉明空间向量中与所述查询向量的距离满足预设条件的汉明空间向量对应的多媒体文件。

[0013] 第二方面，本发明提供一种在网上搜索多媒体文件的装置，包括：

[0014] 获取模块，用于获取与待搜集多媒体文件对应的查询向量；

[0015] 第一划分模块，用于根据预设规则将所述查询向量划分成多个比特块；

[0016] 第一计算模块，针对所述查询向量中的每个比特块，查找与所述比特块对应的索引表确定与所述查询向量对应的数据库向量集中各汉明空间向量中对应比特块和所述查询向量中所述比特块的相似度；所述索引表是根据所述查询向量对应的数据库向量集中各汉明空间向量按照预设规则划分后的对应比特块得到的，所述索引表包括与所述比特块的比特数对应个数的表项，每个表项包括与所述表项的二进制序号相同的对应比特块所属汉明空间向量的标号；

[0017] 第二计算模块，用于针对每个汉明空间向量，将所述汉明空间向量中各比特块的相似度累加，得到所述汉明空间向量的相似度；

[0018] 第三计算模块，用于计算相似度超过预设阈值的各汉明空间向量与所述查询向量的距离；

[0019] 返回模块，用于返回所述相似度超过预设阈值的汉明空间向量中与所述查询向量的距离满足预设条件的汉明空间向量对应的多媒体文件。

[0020] 本发明实施例提供的在网上搜索多媒体文件的方法和装置，将查询向量分成多个比特块后，根据对应的各比特块之间的相似度，确定数据库向量集中汉明空间向量的相似度，从而仅计算相似度超过预设阈值的各汉明空间向量与查询向量的距离并且返回所述相似度超过预设阈值的汉明空间向量中与所述查询向量的距离满足预设条件的汉明空间向量对应的多媒体文件使绝大多数检索的目标向量被包含在所述相似度超过预设阈值的汉明空间向量中，保证了检索的正确率；而且无需在整个数据库向量中对所有汉明空间向量进行遍历计算，降低了计算的复杂度，减轻了计算对系统资源的占用，可在短时间内在大规模数据库中检索出用户所需的多媒体文件，提高了检索效率。

附图说明

[0021] 图1为本发明中在网上搜索多媒体文件的方法实施例的流程图；

[0022] 图2为本发明在网上搜索多媒体文件的方法实施例中索引表的一种示意图；

[0023] 图3为本发明在网上搜索多媒体文件的方法实施例中索引表的又一种示意图；

[0024] 图4为本发明中在网上搜索多媒体文件的装置实施例的结构示意图。

具体实施方式

[0025] 为使本发明实施例的目的、技术方案和优点更加清楚，下面将结合本发明实施例中的附图，对本发明实施例中的技术方案进行清楚、完整地描述，显然，所描述的实施例是本发明一部分实施例，而不是全部的实施例。基于本发明中的实施例，本领域普通技术人员在没有作出创造性劳动前提下所获得的所有其他实施例，都属于本发明保护的范围。

[0026] 本发明各实施例中所述的多媒体(Multimedia)文件，包括但不限于：文字、图片、照片、声音(包含音乐、语音旁白、特殊音效)、动画和影片，以及通过各种编程语言编写的程式所提供的具有互动功能的事物。

[0027] 图像的视觉特性包括但不限于：图像的颜色、形状、纹理、空间关系等。

[0028] 所述汉明空间向量是每个维度取值为0或1的比特串，用于表征多媒体文件的特性，汉明空间向量间的距离用汉明距离来度量。

[0029] 两个等长汉明空间向量的距离计算方法：在两个汉明空间向量的比特串中对应位置不相同的比特个数就表示两个汉明空间向量的距离。两个汉明空间向量的组成越相近，那么这两个汉明空间向量的距离越相近；比如汉明空间向量0001和0011的距离是1；汉明空间向量1001和0111的距离是3。

[0030] 所述汉明空间向量的k近邻查询的定义为：给定一个查询向量，从包含多个汉明空间向量的数据库向量集中查找出与所述查询向量汉明距离最近的k个汉明空间向量。

[0031] 所述查询向量是指能够表示被查询的多媒体文件特性的汉明空间向量。

[0032] 所述数据库向量集是指：搜集到的能够表征多媒体文件的所有汉明空间向量的集合，表示成$B=\{B_1,B_2,\cdots,B_n\}$。

[0033] 所述相似度$S_1,S_2\cdots S_n$，表示所述数据库向量集中的各汉明空间向量$B_1,B_2,\cdots,B_n$分别与所述查询向量的相似性；如果某个汉明空间向量B_i的相似度S_i越高，说明该汉明空间向量B_i与所述查询向量的相似程度越高。

[0034] 在网上搜索多媒体文件的方法实施例的流程如图1所示，所述方法包括如下步骤：

[0035] 步骤100：获取与待搜集多媒体文件对应的查询向量；

[0036] 其中，所述查询向量是可以表征所述待搜集多媒体文件的汉明空间向量。

[0037] 进一步地，在步骤100之前包括如下步骤：

[0038] a).根据预设规则对所述数据库向量集中各汉明空间向量进行划分；

[0039] 具体地，所述预设规则就是将数据库向量集中的各个汉明空间向量以相同的划分方式划分成多个比特块；

[0040] 把所述数据库向量集中的各个汉明空间向量划分成若干段，每段是汉明空间向量的一个连续的比特字串。划分方式包括但不限于：均匀的划分方式和不均匀的划分方式。划分后的每个段就是步骤a中的所述比特块；在实施例中优选均匀的划分方式。

[0041] 所述均匀的划分方式，是以n等分的形式对所述汉明空间向量进行划分($n\geq 2$)；所述不均匀的划分方式，是除了所述均匀的划分方式之外，其余任意的划分方式。

[0042] 比如汉明空间向量00110101，可采用2等分的划分方式划分成0011和0101二段；也可采用不均匀的划分方式将00110101划分成0011、01和01三段，划分后的段就是划分后的比特块。

[0043] 将所述数据库向量集中各个向量以相同的划分方式划分是指：比如数据库中包括两个汉明空间向量10111001和00001111，如果采用均匀的划分方式，10111001可以被划分成1011和1001两个比特块，那么00001111也采用这种划分方式划分成0000和1111两个比特块。即所述数据库向量集中每个向量划分后的比特块数和各个比特块的长度应该一致。

[0044] b).根据各汉明空间向量划分得到的对应顺序的各比特块，建立对应的索引表，所述对应的索引表包括2^{di}个表项，其中di为对应的每个比特块的比特数，第i个表项包括与i相同的对应比特块所属汉明空间向量的标号；

[0045] 其中，所述步骤b包括如下步骤：

[0046] 1)、创建数量与所述汉明空间向量划分后的比特块数量相同的索引表；

[0047] 具体地，设所述数据库向量集包含n个汉明空间向量，分别表示为$B_1,B_2,\cdots,B_n$，每

个汉明空间向量被划分成m个比特块，则创建m个索引表，分别表示为$Index_1$、$Index_2$、…$Index_m$。

[0048] 2)、每个索引表包括多个表项，每个表项的表项序号用比特串表示，各索引表的表项个数与汉明空间向量中的对应比特块的长度相关，可选地，所述索引表中各表项按照对应表项序号从小到大的顺序排列；

[0049] 其中，设所述汉明空间向量的第i个比特块长度，即比特位数为di，那么索引表$Index_i$的表项个数设定为2^{di}个。

[0050] 3)、将各比特块对应的所述汉明空间向量的标号存入对应的索引表中表项序号与所述比特块相同的表项中，由此获得各汉明空间向量比特块的索引。

[0051] 例如，数据库向量集中的一个汉明空间向量B_u=10011101，划分的方式为：

[0052] m=3，d_1=2，d_2=3，d_3=3。按照划分方式可得：$B_u^1=10$，$B_u^2=011$，$B_u^3=101$。那么索引表总共创建3个，分别是$Index_1$、$Index_2$和$Index_3$。其中$Index_1$的表项个数为2^2=4，$Index_2$和$Index_3$的表项个数均为2^3=8；索引表中的各表项初始化为空，即不包含任何内容。然后将向量B_u的标号u分别存入$Index_1$的表项序号为10的表项(简称$Index_1$的表项10)、$Index_2$的表项011和$Index_3$的表项101中。其中，$Index_1$的表项10表示为index[1][10]，$Index_2$的表项011表示为index[2][011]，$Index_3$的表项101表示为index[3][101]；3个索引表如图2所示。

[0053] 步骤101：根据预设规则将所述查询向量划分成多个比特块；

[0054] 具体地，按照所述数据库向量集中各汉明空间向量的划分方式来划分所述查询向量，得到查询向量的m个比特块，表示成q_1、q_2、…q_m。

[0055] 步骤102：针对所述查询向量中的每个比特块，查找与所述比特块对应的索引表确定与所述查询向量对应的数据库向量集中各汉明空间向量中对应比特块和所述查询向量中所述比特块的相似度；所述索引表是根据所述查询向量对应的数据库向量集中各汉明空间向量按照预设规则划分后的对应比特块得到的，所述索引表包括与所述比特块的比特数对应个数的表项，每个表项包括与所述表项的二进制序号相同的对应比特块所属汉明空间向量的标号。

[0056] 这里的与所述查询向量对应的数据库向量集，是指对应的多媒体类型与所述查询向量表征的多媒体类型相同的数据库向量集。

[0057] 进一步地，步骤102具体包括：

[0058] 分别确定与所述比特块的距离为r的表项序号，r为大于等于0不大于di的整数；

[0059] 根据所述距离r，以及所述距离r对应的相似性因子，得到所述表项序号指向的表项对应的汉明空间向量中对应比特块和所述查询向量中所述比特块的相似度，所述距离r对应的相似性因子与所述距离r成反比。

[0060] 优选的，所述r依次取0至x之间的每个值，x为一小于di的预设值。

[0061] 具体地，对于所述查询向量的比特块q_i，在索引表$Index_i$中找出表项序号与所述q_i存在r比特不相同的所有表项构成的表项集合$\varphi(q_i,r)$；其中，r初值为0，i初值为1，x通常预设为3，即r为[0,3]间的整数；

[0062] 遍历所述$\varphi(q_i,r)$中的各表项，获取存储在各表项中的汉明空间向量的标号；

[0063] 将各表项中序号对应的汉明空间向量中第i个比特块与所述查询向量中第i个比

特块的相似度赋值为与所述距离r对应的相似性因子w_r；

[0064] r=r+1并重复上述步骤直到$r>r_{max}$，其中，r_{max}表示相应比特块的比特数，即di；

[0065] 令i=i+1且r清零，重复以上步骤直到i＞m；其中，m是每个汉明空间向量和查询向量划分得到的比特块的数量。

[0066] 其中，按照如下规则设置w_r：r的值越小，相应的w_r的值越大，比如：r取0，1和2，则相应的$w_0=4$，$w_1=2$，$w_2=1$。

[0067] 进一步地，若一个汉明空间向量中，存在至少一个比特块未得到和所述查询向量中对应比特块的相似度，则将该个汉明空间向量中所述比特块和所述查询向量中对应比特的相似度设置为0。

[0068] 步骤103：针对每个汉明空间向量B_1，B_2，…，B_n，将所述汉明空间向量中各比特块和所述查询向量中对应比特块q_1、q_2、…q_m的相似度累加，得到所述汉明空间向量的相似度S_1，S_2…S_n；

[0069] 步骤104：计算相似度超过预设阈值的各汉明空间向量与所述查询向量的距离；

[0070] 进一步地，所述步骤104包括如下步骤：

[0071] 1).计算相似度超过预设阈值的各汉明空间向量与所述查询向量的距离；

[0072] 2).根据计算结果，按照离所述查询向量由近及远的顺序对各汉明空间向量进行排序，得到所述候选向量集中各汉明空间向量的排序列表；

[0073] 3).从所述排序列表中选出离所述查询向量最近的k个汉明空间向量作为查询结果。这里的k为一预设值。

[0074] 例如：设查询向量为0011，相似度超过预设阈值的4个汉明空间向量分别是{1010，1111，0010，0001}，从这4个中查找与所述查询向量距离最近的2个向量的过程如下：

[0075] 根据两个等长汉明空间向量的距离计算方法可知汉明空间向量1010与所述查询向量的距离为2；汉明空间向量1111与所述查询向量的距离为2；汉明空间向量0010与所述查询向量的距离为1；汉明空间向量0001与所述查询向量的距离为1；

[0076] 根据上述的计算结果，得到如下汉明空间向量排序列表：(0010、0001、1010、1111)；其中，与所述查询向量距离最近的2个向量是汉明空间向量是0010和0001。

[0077] 步骤105：返回所述相似度超过预设阈值的汉明空间向量中与所述查询向量的距离满足预设条件的汉明空间向量对应的多媒体文件。

[0078] 进一步地，所述相似度超过预设阈值的汉明空间向量中与所述查询向量的距离满足预设条件的汉明空间向量有至少两个；所述步骤105包括：根据所述满足预设条件的汉明空间向量对应的多媒体文件形成查询结果列表，所述查询结果列表中多媒体文件按照对应汉明空间向量与所述查询向量的距离从小到大的顺序排列；返回所述查询结果列表。

[0079] 这里的预设条件具体可以是与所述查询向量的距离按照从小到大的顺序排在前k位。

[0080] 例如：相似度超过预设阈值的各汉明空间向量按照相似度从大到小的排序列表为(B_3、B_4、B_1、B_2、B_5)；若k=3，则将所述汉明空间向量B_1、B_3和B_4对应的多媒体文件v_1、v_3和v_4按照对应汉明空间向量与所述查询向量的距离从小到大的顺序(v_3、v_4、v_1)返回给所述用户。

[0081] 通过又一实施例对在网上搜索多媒体文件的方法作进一步描述。

[0082] 根据预设规则对所述数据库向量集中各汉明空间向量进行划分；设数据库向量集

B包括分别表征多媒体文件V_1、V_2、V_3和V_4的4个汉明空间向量：B_1=010100，B_2=010011，B_3=110100，B_4=001101。

[0083] 各汉明空间向量平均划分成两个比特块，每比特块包括3个比特。

[0084] 那么建立2个索引表，$Index_1$，$Index_2$，每个索引表分别包括表项序号为000到111的8个表项；其中，表项Index[1][010]存储汉明空间向量B_1和B_2的标号1和2；表项Index[1][110]存储汉明空间向量B_3的标号3；表项Index[1][001]存储汉明空间向量B_4的标号4；表项Index[2][100]存储汉明空间向量B_1和B_3的标号1和3；表项Index[2][011]存储汉明空间向量B_2的标号2；表项Index[2][101]存储汉明空间向量B_4的标号4；2个索引表如图3所示。

[0085] 对于查询向量Q=001100，将所述Q划分成两个向量块q_1=001，q_2=100。

[0086] 将B_1，B_2，B_3，B_4的相似度s_1、s_2、s_3和s_4初始化为0。

[0087] 设r_{max}=1，即r的取值为0和1，对应的w_0=4，w_1=1。

[0088] 设r=0，x=1，则在$Index_x$，即$Index_1$中，对于查询向量Q的第一个向量块q_1=001：与q_1距离r=0的只有表项Index[1][001]，那么表项集合$\varphi(q_1,0)$只包含表项Index[1][001]；则给所述Index[1][001]中存储的标号4对应的汉明空间向量B_4的相似度增加w_0；

[0089] 令r=r+1；

[0090] 与q_1距离r=1的表项有Index[1][101]、Index[1][011]和Index[1][000]，那么表项集合$\varphi(q_1,1)$包含：Index[1][101]、Index[1][011]和Index[1][000]三个表项；分别给所述Index[1][101]、所述Index[1][011]、所述Index[1][000]中存储的标号对应的汉明空间向量的相似度增加w_1，但由于索引表对应的这些表项均没有存储任何汉明空间向量的标号，所以此时各汉明空间向量的相似度不变；

[0091] 此时r=r+1=2＞r_{max}，令x=x+1且r清零；

[0092] 在所述$Index_2$中，查询向量Q的第二个向量块q_2=100：与q_2距离为0的表项只有表项index[2][100]那么表项集合$\varphi(q_2,0)$只包含表项Index[2][100]；则给所述index[2][100]中存储的标号1和3对应的汉明空间向量B_1和B_3的相似度分别增加w_0；

[0093] r=r+1；

[0094] 与q_2距离为1的表项包括：Index[2][000]、Index[2][110]和Index[2][101]，那么表项集合$\varphi(q_2,1)$包含表项：Index[2][000]、Index[2][110]和Index[2][101]三个表项；分别给所述Index[2][000]、所述Index[2][110]和所述Index[2][101]中存储的标号对应的汉明空间向量的相似度增加w_1，其中，只有index[2][101]中存有标号4，所以仅对汉明空间向量B_4的相似度增加w_1；

[0095] r=r+1=2＞r_{max}；

[0096] 令x=x+1=3＞m=2，则循环结束，得到各汉明空间向量的相似度。具体地，相似度S_1=4，S_2=0，S_3=4，S_4=5。预设阈值为3，那么选择相似度大于3的汉明空间向量构成候选向量集{B_1，B_3，B_4}。

[0097] 设置k=2；

[0098] 计算所述候选向量集中各汉明空间向量与所述查询向量的距离；

[0099] 具体地，所述查询向量Q的比特串为001100，所述B_1的比特串为010100，经过计算所述B_1与所述Q的距离为2；所述B_3的比特串为110100，经过计算所述B_3与所述Q的距离为3；

所述B_4的比特串为001101，经过计算所述B_4与所述Q的距离为1。

[0100] 根据上述的计算结果，得到如下汉明空间向量排序列表：(B_4、B_1、B_3)

[0101] 与所述查询向量距离最近的2个汉明空间向量的是：(B_4、B_1)。

[0102] 将汉明空间向量B_4和B_1对应的所述多媒体文件V_4和V_1以如下的顺序(V_4、V_1)返回给所述用户。

[0103] 基于上述描述，本发明实施例提供的在网上搜索多媒体文件的方法，将查询向量分成多个比特块后，根据对应的各比特块之间的相似度，确定数据库向量集中汉明空间向量的相似度，从而仅计算相似度超过预设阈值的各汉明空间向量与查询向量的距离并且返回所述相似度超过预设阈值的汉明空间向量中与所述查询向量的距离满足预设条件的汉明空间向量对应的多媒体文件使绝大多数检索的目标向量被包含在所述相似度超过预设阈值的汉明空间向量中，保证了检索的正确率；而且无需在整个数据库向量中对所有汉明空间向量进行遍历计算，降低了计算的复杂度，减轻了计算对系统资源的占用，可在短时间内在大规模数据库中检索出用户所需的多媒体文件，提高了检索效率。

[0104] 本领域普通技术人员可以理解：实现上述各方法实施例的全部或部分步骤可以通过程序指令相关的硬件来完成。前述的程序可以存储于一计算机可读取存储介质中。该程序在执行时，执行包括上述各方法实施例的步骤；而前述的存储介质包括：ROM、RAM、磁碟或者光盘等各种可以存储程序代码的介质。

[0105] 图4为本发明在网上搜索多媒体文件的装置的实施例结构示意图，如图4所示，所述装置包括：

[0106] 获取模块30，用于获取与待搜集多媒体文件对应的查询向量；

[0107] 第一划分模块40，用于根据预设规则将所述查询向量划分成多个比特块；

[0108] 第一计算模块50，针对所述查询向量中的每个比特块，查找与所述比特块对应的索引表确定与所述查询向量对应的数据库向量集中各汉明空间向量中对应比特块和所述查询向量中所述比特块的相似度；所述索引表是根据所述查询向量对应的数据库向量集中各汉明空间向量按照预设规则划分后的对应比特块得到的，所述索引表包括与所述比特块的比特数对应个数的表项，每个表项包括与所述表项的二进制序号相同的对应比特块所属汉明空间向量的标号；

[0109] 第二计算模块60，用于针对每个汉明空间向量，将所述汉明空间向量中各比特块的相似度累加，得到所述汉明空间向量的相似度；

[0110] 第三计算模块70，用于计算相似度超过预设阈值的各汉明空间向量与所述查询向量的距离；

[0111] 返回模块80，用于返回所述相似度超过预设阈值的汉明空间向量中与所述查询向量的距离满足预设条件的汉明空间向量对应的多媒体文件。

[0112] 所述装置还包括：

[0113] 第二划分模块，用于根据预设规则对所述数据库向量集中各汉明空间向量进行划分；

[0114] 索引表建立模块，用于根据各汉明空间向量划分得到的对应顺序的各比特块，建立对应的索引表，所述对应的索引表包括2^{di}个表项，其中di为对应的每个所述比特块的比特数，二进制序号为j的表项包括所述各比特块中与j相同的对应比特块所属汉明空间向量

的标号。

[0115] 进一步地，所述第一计算模块50包括：

[0116] 确定单元，用于分别确定与所述比特块的距离为r的表项序号，r为大于等于0不大于di的整数；

[0117] 评分单元，用于根据所述距离r，以及所述距离r对应的相似性因子，得到所述表项序号指向的表项对应的汉明空间向量中对应比特块和所述查询向量中所述比特块的相似度，所述距离r对应的相似性因子与所述距离r成反比。

[0118] 所述r依次取0至x之间的每个值，x为一小于di的预设值；所述第一计算模块50还包括：

[0119] 设置单元，用于若一个汉明空间向量中，存在至少一个比特块未得到和所述查询向量中对应比特块的相似度，则将所述一个汉明空间向量中所述比特块和所述查询向量中对应比特的相似度设置为0。

[0120] 所述返回模块80，具体用于：

[0121] 根据所述满足预设条件的汉明空间向量对应的多媒体文件形成查询列表，所述查询结果列表中多媒体文件按照对应汉明空间向量与所述查询向量的距离从小到大的顺序排列；

[0122] 返回所述查询结果列表。

[0123] 基于上述描述，本发明实施例提供的在网上搜索多媒体文件的装置，在网上搜索多媒体文件的方法和装置，将查询向量分成多个比特块后，根据对应的各比特块之间的相似度，确定数据库向量集中汉明空间向量的相似度，从而仅计算相似度超过预设阈值的各汉明空间向量与查询向量的距离并且返回所述相似度超过预设阈值的汉明空间向量中与所述查询向量的距离满足预设条件的汉明空间向量对应的多媒体文件使绝大多数检索的目标向量被包含在所述相似度超过预设阈值的汉明空间向量中，保证了检索的正确率；而且无需在整个数据库向量中对所有汉明空间向量进行遍历计算，降低了计算的复杂度，减轻了计算对系统资源的占用，可在短时间内在大规模数据库中检索出用户所需的多媒体文件，提高了检索效率。

[0124] 最后应说明的是：以上各实施例仅用以说明本发明的技术方案，而非对其限制；尽管参照前述各实施例对本发明进行了详细的说明，本领域的普通技术人员应当理解：其依然可以对前述各实施例所记载的技术方案进行修改，或者对其中部分或者全部技术特征进行等同替换；而这些修改或者替换，并不使相应技术方案的本质脱离本发明各实施例技术方案的范围。

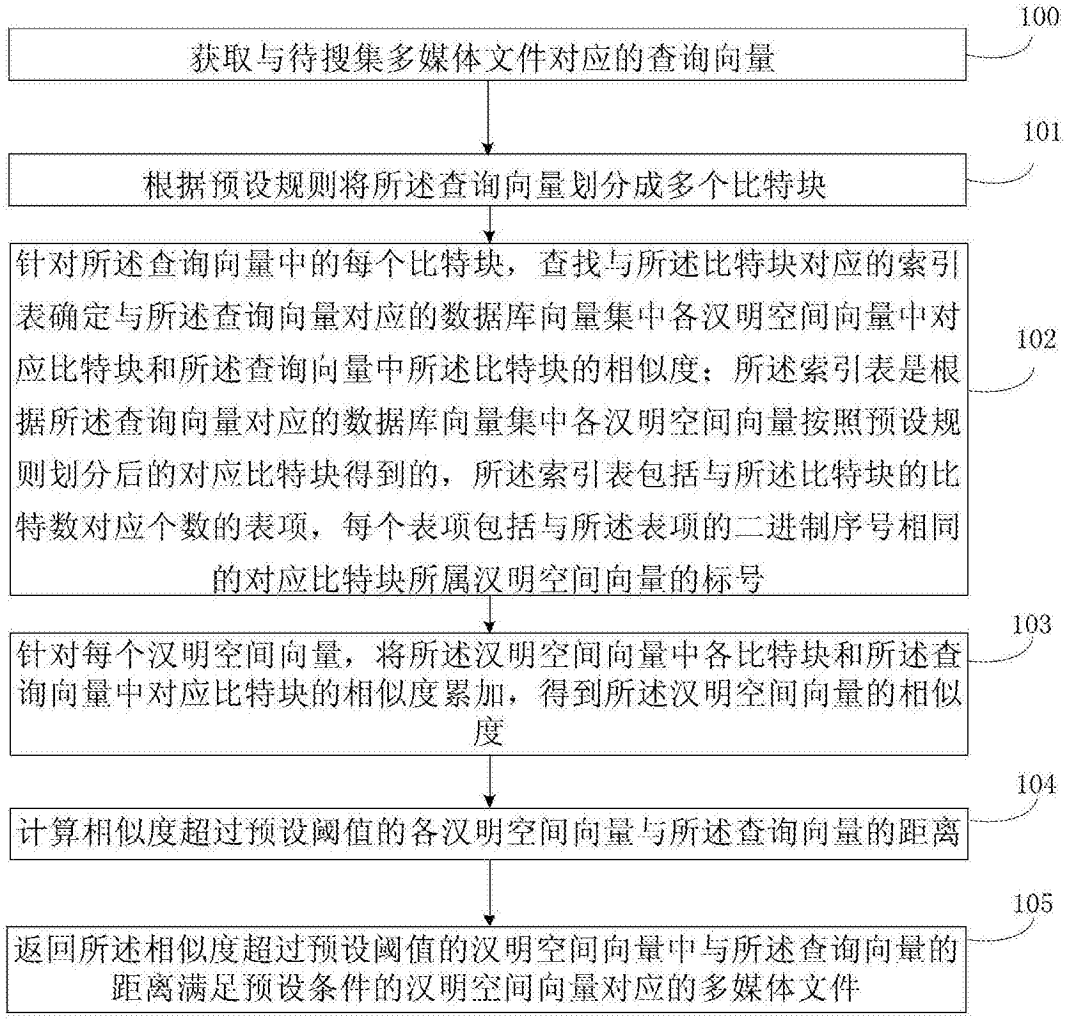

100
获取与待搜集多媒体文件对应的查询向量
101
根据预设规则将所述查询向量划分成多个比特块
102
针对所述查询向量中的每个比特块，查找与所述比特块对应的索引表确定与所述查询向量对应的数据库向量集中各汉明空间向量中对应比特块和所述查询向量中所述比特块的相似度；所述索引表是根据所述查询向量对应的数据库向量集中各汉明空间向量按照预设规则划分后的对应比特块得到的，所述索引表包括与所述比特块的比特数对应个数的表项，每个表项包括与所述表项的二进制序号相同的对应比特块所属汉明空间向量的标号
103
针对每个汉明空间向量，将所述汉明空间向量中各比特块和所述查询向量中对应比特块的相似度累加，得到所述汉明空间向量的相似度
104
计算相似度超过预设阈值的各汉明空间向量与所述查询向量的距离
105
返回所述相似度超过预设阈值的汉明空间向量中与所述查询向量的距离满足预设条件的汉明空间向量对应的多媒体文件

图1

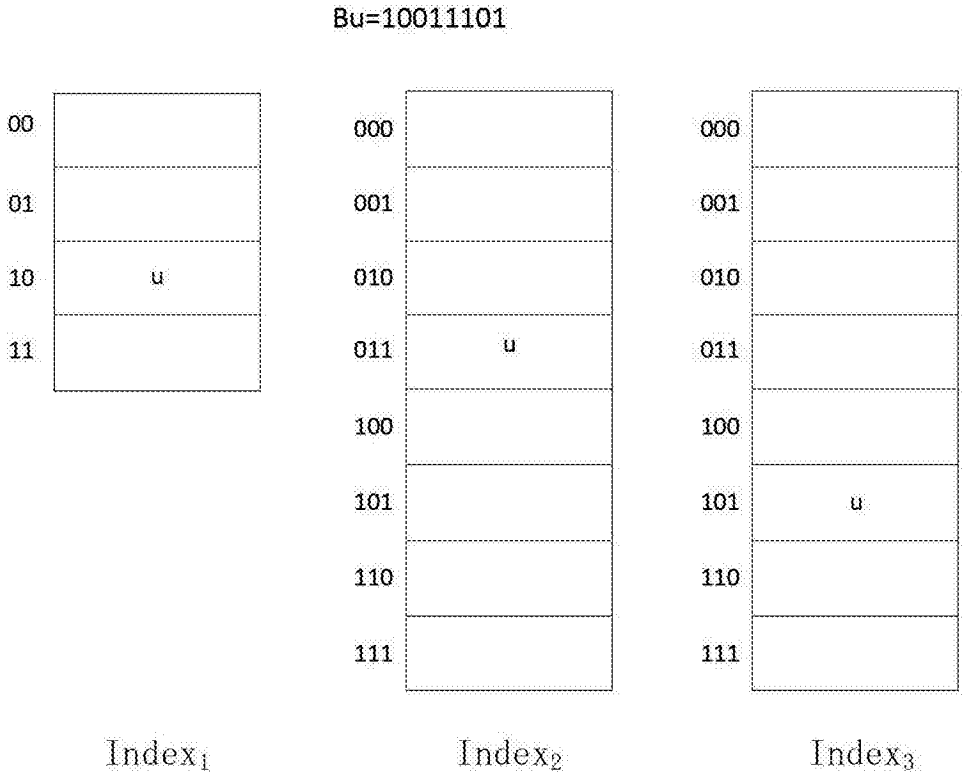

Bu=10011101
00
01
10
u
11
000
001
010
011
u
100
101
110
111
000
001
010
011
100
101
u
110
111
Index1
Index2
Index3

图2

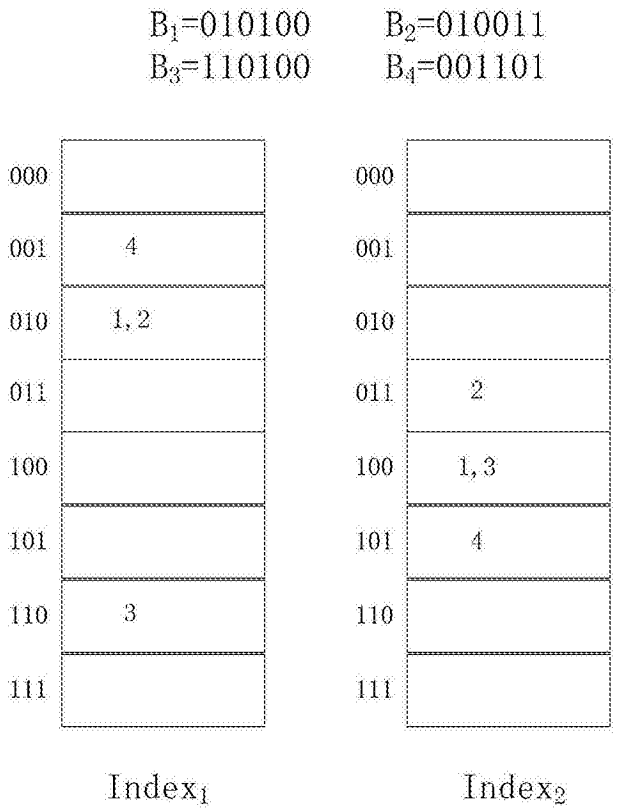

B_1=010100
B_2=010011
B_3=110100
B_4=001101
000
001
4
010
1,2
011
100
101
110
3
111
$Index_1$
000
001
010
011
2
100
1,3
101
4
110
111
$Index_2$

图3

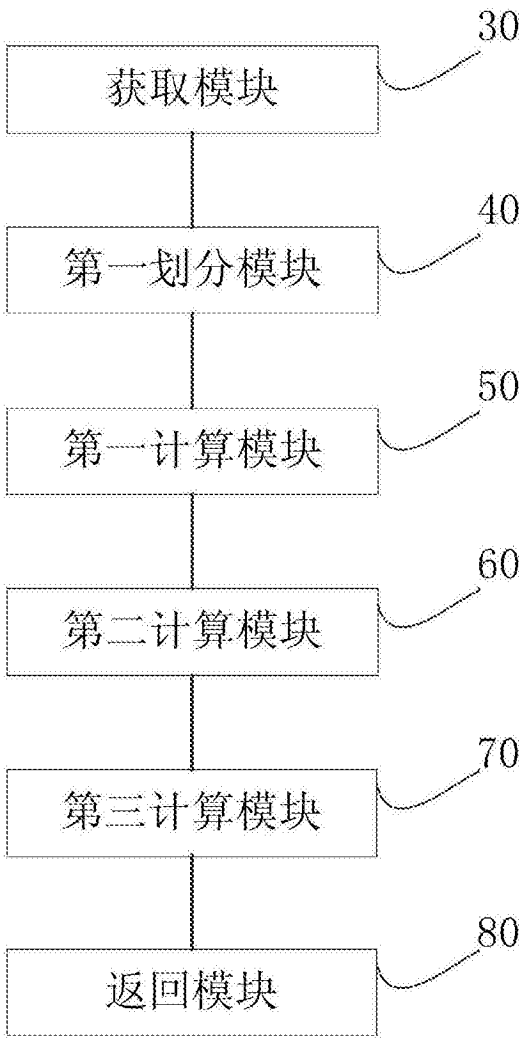

30
获取模块
40
第一划分模块
50
第一计算模块
60
第二计算模块
70
第三计算模块
80
返回模块

图4