(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 104391908 A
(43)申请公布日 2015.03.04

(21)申请号 201410655506.6

(22)申请日 2014.11.17

(71)申请人 南京邮电大学
地址 210003 江苏省南京市鼓楼区新模范马路66号

(72)发明人 韩京宇 陈可佳 曾建辉

(74)专利代理机构 南京知识律师事务所 32207
代理人 汪旭东

(51)Int.Cl.
G06F 17/30(2006.01)

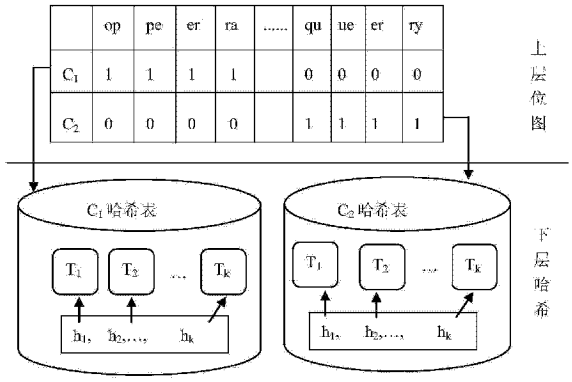
权利要求书1页 说明书4页 附图2页

(54)发明名称

一种图上基于局部敏感哈希的多关键字索引方法

(57)摘要

本发明公开了一种图上基于局部敏感哈希的多关键字索引方法，属于图数据(graph data)管理技术领域，该方法是双层索引来支持图上的多关键字查询。若干图根据顶点关键字在n-gram空间聚类后，根据聚簇结构构建上层的位图和下层的局部敏感哈希表：上层的位图根据关键字包含的粗粒度的n-gram(n个连续字母构成的字符串)实现图到类簇的映射；下层的每个类簇对应一个局部敏感哈希表，哈希表的桶中包含细粒度的n-gram对应的候选图。该索引具有如下优势：(1)查询I/O和关键字个数独立，显著减少多关键字查询的I/O次数，加快查询速度；(2)不同粒度的n-gram相结合，有效避免索引对拼写错误敏感，提高了概率返回期望的结果。



CN 104391908 A

1. 一种图上基于局部敏感哈希的多关键字索引方法，其特征在于：所述方法的图经过聚类后，每个类簇用一个粗粒度的 n-gram 位串来表征，查询时，根据关键字的粗粒度位串和类簇位图的匹配结果识别候选类簇，包括如下步骤：

步骤 1：类簇位图表示；

根据图顶点包含的关键字，将所有图映射到一个粗粒度的 n-gram 空间，如果 n-gram 空间共有 N 个不同的 n-gram，每个类簇对应一个长度为 n 的位串，类簇对应的关键字如果包含第 i 个 n-gram，则位串对应的位为 1，否则为 0，所有类簇的位串构成位图，记为 BT；

步骤 2：查询位串表示；

根据上述步骤 1 的 n-gram 空间，构建多关键字查询位串 Q，查询的关键字包含某个 n-gram，则查询位串对应的位为 1，否则为 0；

步骤 3：类簇匹配；

如果位图 BT 中某个位串恰好涵盖查询位串 Q 中所有为 1 的位，则该位串的对应类簇是候选类簇。

2. 根据权利要求 1 所述的一种图上基于局部敏感哈希的多关键字索引方法，其特征在于，所述方法的磁盘访问次数独立于关键字个数，减少磁盘 I/O；使用所述索引包括如下步骤：

步骤 1：哈希索引构建；在一个类簇中，给定图的所有关键字组合 $\{CM_1, \cdots, CM_i, \cdots, CM_n\}$，一个关键字组合 $CM_i = (w^1, \cdots, w^j, \cdots, w^m)$，wj 代表关键字，设 CM_i 对应的细粒度 n-gram 集合记为 $NG(CMi) = (g^1, \cdots, g^j, \cdots, g^M)$，$g^j$ 是一个细粒度的 n-gram；给定一个局部敏感哈希函数族的 k 个哈希函数 $\{h_1, \cdots, h_j, \cdots, h_k\}$，每个哈希函数 h_j 对应一个哈希表 T_j，每个哈希函数 h_j 作用于 NG(CMi) 的 M 个 n-gram 上，计算出哈希值，将 CM_i 对应的图存储到 T_J 相应桶中；

步骤 2：查询关键字匹配；在细粒度的 n-gram 空间，查询 Q 的多个关键字表征为 $NG(Q) = \{q^1, \cdots, q^j, \cdots, q^n\}$，$q^j$ 代表一个细粒度的 n-gram，根据步骤 1 的 k 个哈希函数 $\{h_1, \cdots, h_j, \cdots, h_k\}$，将 Q 分别映射到 k 张哈希表 $T_1, \cdots, T_j, \cdots, T_k$ 上，在冲突的桶上获取匹配图。

3. 根据权利要求 1 所述的一种图上基于局部敏感哈希的多关键字索引方法，其特征在于：所述方法支持图关键字查询，基于粗粒度 n-gram，即：n 个连续字母构成的字符串的位图索引。

4. 根据权利要求 1 所述的一种图上基于局部敏感哈希的多关键字索引方法，其特征在于：所述方法应用于关联复杂数据，即：链接的 Web 网页、社会网络、蛋白质交互网络的图结构数据的存储和查询。

一种图上基于局部敏感哈希的多关键字索引方法

技术领域

[0001] 本发明涉及一种图上基于局部敏感哈希的多关键字索引方法，属于图数据管理技术领域。

背景技术

[0002] 近年来，万维网、社会网络、生物医学、化合物分子结构等领域积累了大量关联复杂的数据，其结构通常抽象为图。为了实现信息获取和分析，图上的关键字查询是一个基本问题。目前图上关键字查询存在以下问题：(1) 由于图数据关系复杂，查询分析时经常涉及多个关键字。目前图关键字索引主要基于倒排表及其变体，要依次读取每个（对）关键字对应的候选图，I/O 随着关键字个数线性增长，查询效率不高；(2) 图数据的索引常假定数据及查询精确且完整，实际无论是图数据还是用户查询，经常会有个别字母拼写错误，这时无法利用索引得到查询结果。

[0003] 有两种支持图关键字查询的方案，一种方案将图的顶点和边作为记录存储到关系数据库，利用多表连接获得查询结果。Banks 是第 28 届国际大数据库会议 (International Conference on Very Large Data Bases, VLDB) 上《BANKS:Browsing and keyword searching in relational databases》提出的系统，该系统利用关系数据库的多表连接实现关键字检索和浏览。DBXplorer 是第 18 届国际数据工程 (International Conference on Data Engineering, ICDE) 会议《DBXplorer:A system for keyword-based search over relational databases》提出的支持图关键字查询的系统，该系统提供图形界面，用户交互式地对图结构数据进行关键字查询。另外一种方案直接在图上建索引实现关键字查询支持。EASE 是《Information Systems》杂志第 36 卷第 2 期提出的一种支持图结构数据的通用索引，它分别将关系表、文档和 XML 元素当作图的顶点，将主外键、IDERF 和超链接当作图的边，构建倒排表支持关键字查询。2007 年 SIGMOD 会议上，《BLINKS:Ranked Keyword Searches on Graphs》一文提出 Blinks 双层索引，通过块索引 (block index) 和块内索引 (intra-block index) 支持图上的关键字查询。专利《基于 Datalog 的分布式环境下大图数据查询方法》（申请号 CN201210210245，公开号 CN102799624A) 的发明设计了一种基于 Datalog 的分布式环境下大图数据查询方法，它利用等价规则和统计数据，提高大图查询执行计划的效率。专利《一种基于邻接节点树的网络图索引方法》（申请号：CN201210063543，公开号：CN102662974B) 公开了一种在大型网络图中基于邻接节点树的子图查询方法，该发明利用节点间的邻接关系构建索引树，并以邻接节点树作为大型网络图的索引特征实现子图查询。但这些工作未涉及如何减少多关键字查询 I/O 的问题。

[0004] 局部敏感哈希 (Locality Sensitive Hashing, LSH) 是一种解决近似查询的有效技术，其基本思想是把相似的数据对象以高的概率哈希到相同桶中，非相似的数据对象几乎不被哈希到相同的桶中。查询时，把冲突桶中的数据对象作为候选集。针对不同应用需求和相似性度量，目前提出了多种哈希技术。针对高维数据的余弦相似度计算，文献《Compact projection:Simple and efficient near neighbor search with practical

memory requirements》(Proc.of the IEEE Computer Society Conf.on Computer Vision and Pattern Recognition,San Francisco:IEEE Computer Society,2010:3477-3484)提出了随机投影敏感哈希函数族。对基于集合的雅可比相似度,文献《On the resemblance and containment of documents》(Proc.of the 1997 International Conference on Compression and Complexity of Sequences.Positano:IEEE Computer Society,1997:21-29)提出了最小哈希(Min-Hash)技术,用于近似计算集合数据的相似度。专利《基于加权编辑距离的模糊关键字查询方法及系统》(申请号:CN201010240402,公开号:CN101916263B)设计了一种基于加权编辑距离的模糊关键字查询方法,可以更有效地返回用户真实查询的数据,提高用户的满意度。但目前缺少如何应用局部敏感哈希解决图上的多关键字查询的工作。

[0005] 以上两个方面的工作没有涉及如何有效地支持多关键字查询,减少磁盘 I/O 的问题,也没有涉及查询时对关键字拼写错误过分敏感的问题。而本发明能够很好地解决上面的问题。

发明内容

[0006] 本发明目的在于提供了一种图上基于局部敏感哈希的多关键字索引方法,该方法支持图上的多关键字查询,并且该方法解决了关键字容错、位图和局部敏感哈希相结合的双层索引问题。

[0007] 本发明解决其技术问题所采取的技术方案是:本发明提供了一种图上基于局部敏感哈希的多关键字索引方法,该方法支持图关键字查询的、基于粗粒度 n-gram(即:n 个连续字母构成的字符串)的位图索引,图经过聚类后,每个类簇用一个粗粒度的 n-gram 位串来表征,查询时,根据关键字的粗粒度位串和类簇位图的匹配结果识别候选类簇,包括以下三个步骤:

[0008] 步骤 1:类簇位图表示;根据图顶点包含的关键字,将所有图映射到一个粗粒度的 n-gram 空间,如果 n-gram 空间共有 N 个不同的 n-gram,每个类簇对应一个长度为 n 的位串,类簇对应的关键字如果包含第 i 个 n-gram,则位串对应的位为 1,否则为 0,所有类簇的位串构成位图,记为 BT;

[0009] 步骤 2:查询位串表示;根据步骤 1 的 n-gram 空间,构建多关键字查询位串 Q,查询的关键字包含某个 n-gram,则查询位串对应的位为 1,否则为 0;

[0010] 步骤 3:类簇匹配;如果位图 BT 中某个位串恰好涵盖查询位串 Q 中所有为 1 的位,则该位串的对应类簇是候选类簇。

[0011] 本发明的磁盘访问次数独立于关键字个数,有效减少了磁盘 I/O,使用该索引分成以下两个步骤:

[0012] 步骤 1:哈希索引构建;在一个类簇中,给定图的所有关键字组合 $\{CM_1,\cdots,CM_i,\cdots,CM_n\}$,一个关键字组合 $CM_i=(w^1,\cdots,w^j,\cdots,w^m)$,$w^j$ 代表关键字,设 CM_i 对应的细粒度 n-gram 集合记为 $NG(CMi)=(g^1,\cdots,g^j,\cdots,g^M)$,$g^j$ 是一个细粒度的 n-gram;给定一个局部敏感哈希函数族的 k 个哈希函数 $\{h_1,\cdots,h_j,\cdots,h_k\}$,每个哈希函数 h_j 对应一个哈希表 T_j,每个哈希函数 hj 作用于 NG(CMi) 的 M 个 n-gram 上,计算出哈希值,将 CM_i 对应的图存储到 T_J 相应桶中;

[0013] 步骤 2：查询关键字匹配；在细粒度的 n-gram 空间，查询 Q 的多个关键字表征为 $NG(Q) = \{q^1, \cdots, q^j, \cdots, q^n\}$，$q^j$ 代表一个细粒度的 n-gram，根据步骤 1 的 k 个哈希函数 $\{h_1, \cdots, h_j, \cdots, h_k\}$，将 Q 分别映射到 k 张哈希表 $T_1, \cdots, T_j, \cdots, T_k$ 上，在冲突的桶上获取匹配图。

[0014] 本发明应用于关联复杂数据，如链接的 Web 网页、社会网络、蛋白质交互网络等图结构数据的存储和查询。

[0015] 有益效果：

[0016] 1、本发明查询 I/O 次数和关键字个数独立，避免了 I/O 随关键字个数线性增长，提高了查询效率。

[0017] 2、本发明能够对关键字容错，即便图顶点或查询关键字有拼写错误，仍以高的概率返回候选匹配图。

附图说明

[0018] 图 1 本发明的方法流程图。

[0019] 图 2 为本发明的类簇和对应关键字示例图。

[0020] 图 3 为本发明的双层索引示例图。

[0021] 图 4 查询时的哈希过程示例图。

具体实施方式

[0022] 下面结合说明书附图，对本发明的优选实施例进行详细描述。

[0023] 如图 1 所示，本发明提供了一种图上基于局部敏感哈希的多关键字索引方法，该方法支持图关键字查询的、基于粗粒度 n-gram（即：n 个连续字母构成的字符串）的位图索引，图经过聚类后，每个类簇用一个粗粒度的 n-gram 位串来表征，查询时，根据关键字的粗粒度位串和类簇位图的匹配结果识别候选类簇，包括以下三个步骤：

[0024] 步骤 1：类簇位图表示；根据图顶点包含的关键字，将所有图映射到一个粗粒度的 n-gram 空间，如果 n-gram 空间共有 N 个不同的 n-gram，每个类簇对应一个长度为 n 的位串，类簇对应的关键字如果包含第 i 个 n-gram，则位串对应的位为 1，否则为 0，所有类簇的位串构成位图，记为 BT；

[0025] 步骤 2：查询位串表示；根据步骤 1 的 n-gram 空间，构建多关键字查询位串 Q，查询的关键字包含某个 n-gram，则查询位串对应的位为 1，否则为 0；

[0026] 步骤 3：类簇匹配；如果位图 BT 中某个位串恰好涵盖查询位串 Q 中所有为 1 的位，则该位串的对应类簇是候选类簇。

[0027] 本发明的磁盘访问次数独立于关键字个数，有效减少了磁盘 I/O，使用该索引分成以下两个步骤：

[0028] 步骤 1：哈希索引构建；在一个类簇中，给定图的所有关键字组合 $\{CM_1, \cdots, CM_i, \cdots, CM_n\}$，一个关键字组合 $CM_i = (w^1, \cdots, w^j, \cdots, w^m)$，$w^j$ 代表关键字，设 CM_i 对应的细粒度 n-gram 集合记为 $NG(CMi) = (g^1, \cdots, g^j, \cdots, g^M)$，gj 是一个细粒度的 n-gram；给定一个局部敏感哈希函数族的 k 个哈希函数 $\{h_1, \cdots, h_j, \cdots, h_k\}$，每个哈希函数 h_j 对应一个哈希表 T_j，每个哈希函数 hj 作用于 NG(CMi) 的 M 个 n-gram 上，计算出哈希值，将 CM_i 对应的图存储到 T_J 相应桶

中；

[0029] 步骤 2：查询关键字匹配；在细粒度的 n-gram 空间，查询 Q 的多个关键字表征为 $NG(Q) = \{q^1, \cdots, q^j, \cdots, q^n\}$，$q^j$ 代表一个细粒度的 n-gram，根据步骤 1 的 k 个哈希函数 $\{h_1, \cdots, h_j, \cdots, h_k\}$，将 Q 分别映射到 k 张哈希表 $T_1, \cdots, T_j, \cdots, T_k$ 上，在冲突的桶上获取匹配图。

[0030] 本发明应用于关联复杂数据，如链接的 Web 网页、社会网络、蛋白质交互网络等图结构数据的存储和查询。

[0031] 如图 2 所示，本发明假设共有 6 张图结构数据 G_1、G_2、G_3、G_4、G_5 和 G_6，其包含的关键字分别是 $kw(G_1) = \{operating\ system, process, task\ schedule, real\ time\}$，$kw(G_2) = \{operating\ system, process, batch\ processing\}$，$kw(G_3) = \{database, big\ data, query, cloud\ computing\}$，$kw(G_4) = \{data, query, cloud\ computing\}$，$kw(G_5) = \{graph\ database, data, query, rdf\ data\}$ 和 $kw(G_6) = \{relation\ model, data, query, rdf\ data\}$，被分成两个类簇，$C_1 = \{G_1G_2\}$ 和 $C_2 = \{G_3G_4G_5G_6\}$。

[0032] 根据上述数据，构建如图 3 所示的双层索引：上层是位图索引（本例子以 2-gram 为粗粒度 n-gram），如果 C_1 对应的单元格为 1，表示 C_1 中关键字包含该 2-gram，例如'op','pe','er','ra'等；如果 C_1 对应的单元格为 0，表示 C_1 中关键字不包含该 2-gram，例如'qu','ue','er'和'ry'等；下层是局部敏感哈希索引，假定有 k 个哈希函数，则每个类簇内部有 k 张哈希表，不妨记为 $T_1, \cdots, T_k$。

[0033] 设哈希表取 3-gram 作为细粒度的 n-gram（注意：哈希表的 n-gram 要长于上层位图采用的 n-gram），下面以局部敏感哈希函数 h_1 为例，介绍哈希表 T_1 的构建过程：(1) 将一张图的一个关键字组合的每个关键字取其所有的 3-gram，将这个关键字组合的所有 3-gram 当作一个集合；(2) 对集合中每个 3-gram，计算其模值，模值＝（每个字符对应 ASCII 码的和）mod 10；(3) 取最小模值作为最终的哈希函数值，根据此哈希值将图映射到对应的桶中。据此可以创建第一张哈希表 T_1，如图 4 所示意。

[0034] 给定多关键字查询 $Q = \{query, database, rdf\}$，根据哈希函数 h_1，查询结果的计算过程如下：

[0035] (1) 计算 Q 的所有 3-gram，记为：

[0036] $NG(Q) = \{que, uer, ery, dat, ata, tab, aba, bas, ase, rdf\}$；

[0037] (2) 根据哈希函数 h_1 计算 NG(Q) 中每个 3-gram 的对应模值，即 {7,5,4,3,4,8,9,6,5,6}，可知最小模值是 3，这就是最终的哈希函数值。

[0038] (3) 根据哈希函数值，将图映射到哈希表 T_1 对应的第 4 个桶（桶编号为 3），取候选图 G_5 和 G_6。

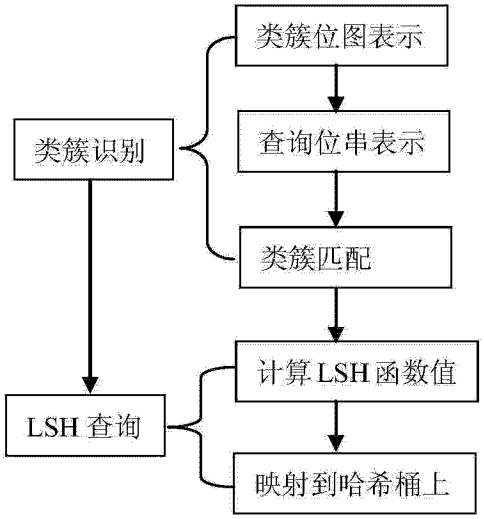


图 1

类簇	图 id	关键字	2-gram 集合示意	3-gram 集合示意
C_1	G_1	{operating system, process, task schedule, real time}	{op,pe,er,ra,at,ti,in,ng,…, re,ea,al,ti,im,me}	{ope,per,era,rat,ati, ...,rea,eal,tim,ime}
	G_2	{operating system, process, batch processing}	{op,pe,er,ra,at,ti,in,ng,…, pr,ro,oc,ce,es,ss,si,in,ng}	{ope,per,era,rat,ati, ...,ces,ess,ssi,sin,ing}
C_2	G_3	{database, big data, query, cloud computing}	{da,at,ta,ab,ba,as,se,bi,…, co,om,mp,pu,ut,ti,in,ng}	{dat,ata,tab,aba,bas,ase ,...,mpu,uti,tin,ing}
	G_4	{data, query, cloud computing}	{da,at,ta,qu,ue,er,ry,…, co,om,mp,pu,ut,ti,in,ng}	{dat,ata,que,uer,..., mpu,uti,tin,ing}
	G_5	{graph database, data, query, rdf data}	{gr,ra,ap,ph,da,at,ta,ab,…, qu,ue,er,ry,rd,df,da,at,ta}	{gra,rap,aph,dat,ata,..., uer,ery,rdf,dat,ata}
	G_6	{relation model, data, query, rdf data}	{re,el,la,at,ti,io,on,mo,,…, qu,ue,er,ry,rd,df,da,at,ta}	{rel,ela,lat,ati,..., uer,ery,dat,ata}

图 2

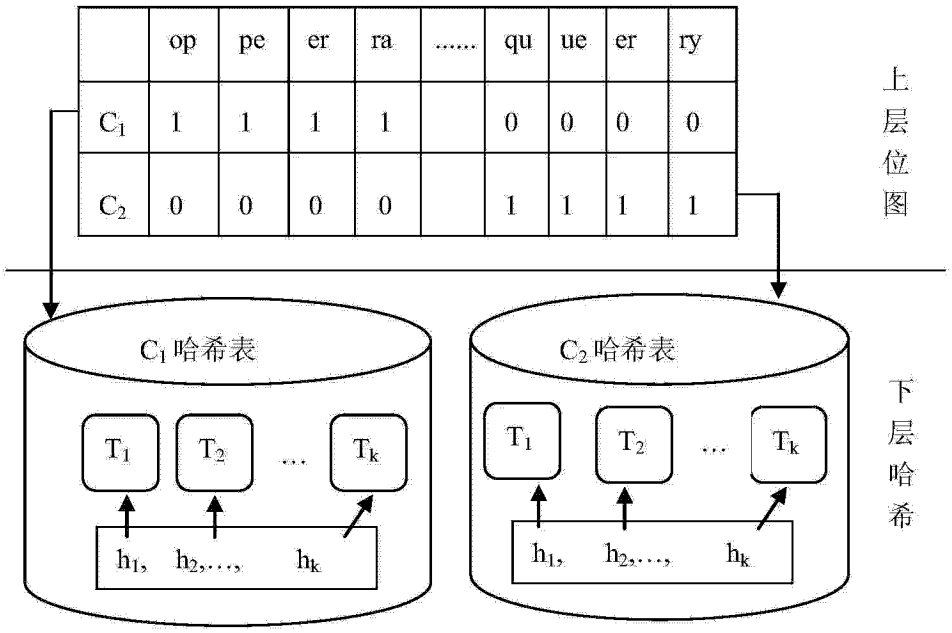


图 3

桶编号（哈希值）	桶中的图 id
0	G_3,G_4
1	-
2	-
3	G_5 G_6
4	-
….	…
9	-

图 4