

基于关联图模型的医学图像 Top- k 查询方法

李鹏远¹ 潘海为¹ 李 青² 韩启龙¹ 谢晓芹¹ 张志强¹

¹(哈尔滨工程大学计算机科学与技术学院 哈尔滨 150001)

²(香港城市大学计算机工程与信息技术系 香港)

(pengyuan@live.cn)

Top- k Medical Images Query Based on Association Graph

Li Pengyuan¹, Pan Haiwei¹, Li Qing², Han Qilong¹, Xie Xiaoqin¹, and Zhang Zhiqiang¹

¹(College of Computer Science and Technology, Harbin Engineering University, Harbin 150001)

²(Department of Computer Engineering and Information Technology, City University of Hong Kong, Hong Kong)

Abstract Patient-to-patient comparison, especially image-to-image comparison plays an important role in the medical domain since doctors invariably make diagnoses based on prior experiences of similar cases. It is very significant for doctors to find similar medical images from the database as similar pathological changes in prior patients' images and corresponding reports can assist doctors to make diagnoses for current patients. Therefore, advanced medical image retrieval techniques have been widely studied to improve the accuracy in recent years. However, the processing time has become another problem in medical image retrieval domain because of the increasing number of medical images. As doctors are only interested in the most similar k results, a novel model of association graph is proposed for medical image top- k query in this paper. The fuzzy expression in a association graph can describe the similarity between images effectively. Moreover, a series of correlation measurements are proposed for similarity reasoning. Then the medical image top- k query method is represented based on the characters of correlation measurements. Furthermore, four walk strategies are studied to accelerate and stabilize the top- k process. Experimental results show that its efficiency and effectiveness are higher in comparison with state of the art.

Key words association graph; Top- k query; walk strategy; image retrieval; medical image

摘 要 找到与病人具有相似纹理特征的医学图像,有助于医生结合历史病历信息对病人作出更为准确的诊断.基于此,大量的研究工作围绕如何提高基于内容的医学图像检索技术的准确性展开.然而,现有的基于内容的医学图像检索技术均是基于查询图像与数据库中图像的逐张匹配过程,面对迅速增长的医学图像数量,查询等待时间过长成为医学图像检索领域的另一主要问题.鉴于用户往往只对前 k (Top- k) 个检索结果感兴趣,提出了一种基于关联图模型的医学图像 Top- k 查询方法.首先,提出一种关联图模型,使用该模型可以有效地刻画医学图像之间关联关系的模糊性;继而利用关联图模型,提出一系列关联性度量计算方法,从而使得仅需对图像匹配一次即可更新所有图像与查询图像之间的相似度范围.由此,提出 Top- k 查询方法以及基于游走的查询优化策略.实验证明提出的方法可以有效地减少图像匹配次数,降低时间复杂度.

收稿日期:2014-07-25;修回日期:2014-11-20

基金项目:国家自然科学基金项目(61272184,61202090,61100007);教育部新世纪优秀人才支持计划项目(NCET-11-0829);中央高校基本科研业务费专项基金项目(HEUCF100602,HEUCFT1202)

通信作者:潘海为(panhaiwei2006@hotmail.com)

关键词 关联图; Top- k 查询; 游走策略; 图像检索; 医学图像

中图法分类号 TP391

随着医学成像技术及相关技术的提升,医学图像(例如 CT 图像、ECT 图像、核磁共振图像)正越来越多地被应用于对病人病情的诊断以及预测等方面,并起到了显著的效果.这一情况导致了医学图像在数量上的飞速增长.据统计,日内瓦大学医学院放射线科 2002 年平均每天产生 12 000 张图像,到 2009 年平均每天产生 114 000 张图像^[1].医学图像应用的普及是因为它为医生提供了一种可能,即通过医学图像便可观察病人身体内部器官、组织的变化情况.然而,医学图像为人类生活带来便利的同时,也同样要求医生在通过医学图像进行诊断的过程中,需要掌握大量的病理知识、经验知识甚至成像知识.拥有不同知识背景的医生即使对同一张医学图像,其诊断结果也往往存在差异.一篇关于“CT 图像导致的诊断错误”的文献^[2]指出,43%的误诊是因为医生没有准确地观察图像,48%是因为医生不能完全理解图像所要表达的信息.目前,医学界为了降低由医学图像导致的误诊率所采用的方法主要包括:1)增加医生的专业知识;2)让医生更多地学习历史病例及临床症状;3)要求医生与历史病例作对比后再作诊断.可见,降低误诊率的关键在于对历史病例的学习与利用.

通过基于内容的医学图像检索技术可以在数据库中找到与患者图像相似的医学图像,从而发现之前和患者患有相同或相似病理情况的病人.通过对以往医生对此类病人作出的决策和病人病情发展情况的学习,有助于医生结合历史病例信息对现有病人作出更为准确的诊断,从而降低误诊率.可见,基于内容的医学图像检索技术具有很强的实用价值和社会价值.

然而,由于医学图像检索技术较一般图像检索技术具有特殊性(如:对图像内容变化敏感、医学本体难以识别与表示等问题),大量的研究工作仍集中在如何提高医学图像检索技术的准确性方面.由于越来越复杂的图像特征表示模型及匹配方法被用于基于内容的医学图像检索技术中,飞速增长的医学图像数量、查询等待时间过长已经成为基于内容的医学图像检索领域的另一主要问题.

通过对医生的调查,我们了解到医生在利用历史病例进行辅助诊断的过程中往往只关心和患者最为相似的 k 个历史病例信息.也就是说,医生仅关心

医学图像检索系统返回的前 k 个最为相似的医学图像信息.然而,现有的基于内容的医学图像检索技术均是基于查询图像与数据库中所有医学图像的逐张匹配过程,随着医学图像数量的不断增长,现有的基于内容的医学图像检索技术将导致用户的查询等待时间越来越长.究其原因,是因为过多的时间被用于匹配数据库中那些不可能和查询图像具有高相似度的图像.如何制定有效的规则来减少查询图像与数据库中图像的匹配次数,从而通过对数据库中一部分图像进行匹配即可得到与查询图像最为相似的前 k 个结果是本文研究的主要内容.据作者所知,目前还没有基于内容的医学图像 Top- k 查询的相关研究.

图模型作为一种有效的描述事物之间关系的数据模型被广泛研究与应用.本文首先基于传统的有向图提出了一种带有模糊性度量的关联图模型,有效地刻画了数据库中图像之间的关联关系;然后,针对这些关联关系,提出了一系列关联性度量计算方法,实现了通过已知关联关系对未知关联关系的预测,使得仅需对图像匹配一次即可更新数据库中所有图像与查询图像的相似度范围;继而,提出了基于关联图模型的医学图像 Top- k 查询方法.为了进一步提高 Top- k 查询方法的高效性与稳定性,本文提出了 3 种基于游走的查询优化策略.

本文的主要贡献如下:

- 1) 提出一种带有模糊性度量的关联图 (association graph) 模型;
- 2) 提出一系列关联性度量计算方法;
- 3) 根据关联性度量方法提出一种基于内容的医学图像 Top- k 查询方法;
- 4) 提出 3 种基于游走的查询优化策略.

1 相关工作

由于医学图像中往往蕴涵着复杂的语义信息,而对于图像的描述总是有限的,所以,相较于基于描述的医学图像检索技术,基于内容的图像检索技术能够更为有效地利用图像本身固有的内容特征进行检索,从而返回更为准确的查询结果.一个完整的基于内容的医学图像检索过程往往包含医学图像预处理与医学图像检索 2 部分.

由于医学图像在生成过程中会同时产生许多附加信息(如病人的姓名、年龄、拍摄时间等信息),而这些附加信息不利于图像内容的相似性判断,所以基于内容的医学图像检索过程通常需要首先对医学图像进行预处理,即去除医学图像中无用的背景信息^[3]. 现有的医学图像预处理技术可以分为 3 类:1) 手动的医学图像预处理技术,如利用 Matlab 软件中的函数 *roipoly()* 手工地对医学图像的感兴趣区域(region of interest, ROI)进行提取;2) 半自动的医学图像预处理技术,如 Wiemker 等人^[4]提出的基于最大化拉普拉斯积分区域的方法;3) 自动的医学图像预处理技术,如 Lehmann 等人^[5]提出的基于 4 种评判标准的自动医学图像预处理技术. 图 1 给出了一种半自动的医学图像预处理过程的实例^[6],我们可以发现医学图像在基于内容的检索中具有背景单一、图像内医学本体大小和位置相对固定等特点.

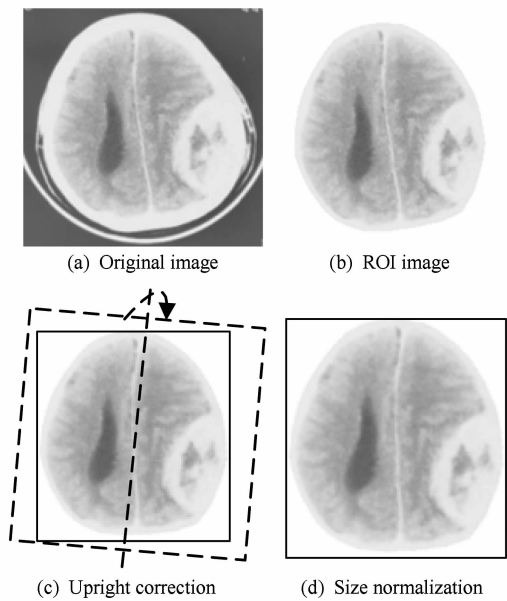


Fig. 1 Medical image preprocess.

图 1 医学图像预处理实例

基于内容的医学图像检索方法是指通过图像的颜色、纹理等特征进行相似性检索. 纹理信息作为医生通过医学图像进行诊断的关键被广泛应用于基于内容的医学图像检索技术中. 传统的方法是通过小波变换、傅里叶变换、Gabor 变换等方法来提取图像的纹理特征^[7],从而进行相似性检索. 然而对于医学图像来说,全局特征往往不能对医学图像纹理特征的局部变化进行有效的表示. 根据对医生的咨询我们了解到,在临床诊断过程中决定诊断结果的图像信息往往只是图像本身的很小一部分. 可见,基于全局图像特征的图像检索技术很难针对医生感兴趣的

区域进行相似性判断. 于是,Unay 等人^[8]提出空间局部二值模式(spatial local binary patterns^[9], Sp-LBP)来表示医学图像的纹理特征,通过对医学图像进行分块得到图像的局部区域特征;文献^[10]结合大量的医学病理知识提出 9 点图模型用以对脊柱轮廓进行表示,从而完成检索过程;潘海为等人^[6]提出使用不确定定点图对医学图像纹理特征进行表示. 这些方法都极大地提高了基于内容的医学图像检索技术的准确性.

然而,目前基于内容的医学图像检索技术均是基于查询图像与数据库中所有图像的匹配过程. 随着医学图像数量的飞速增长,查询等待时间过长已经成为基于内容的医学图像检索领域的另一主要问题. 目前,在文本信息搜索、语义库查询、蛋白质交互网络分析等领域已有大量的关于 Top-*k* 的研究工作^[11-12]. 在一般图像检索领域,图像 Top-*k* 查询技术作为解决当今图像数量爆炸式增长问题的关键被广泛研究,其主要的方法是首先提取图像的特征向量,将图像特征转换为可以进行排序的数值特征进行 Top-*k* 查询. 文献^[13]使用文本信息对图像进行 Top-*k* 查询;文献^[14]使用灰度直方图的索引结构进行 Top-*k* 查询;文献^[15]使用 feedback 机制提高图像检索方法的检索速度;文献^[16]使用混合方法提高图像检索的准确性,降低检索时间. 然而,因为医学图像检索技术所使用图像特征模型具有其特殊性,对于一般图像的 Top-*k* 查询方法不能直接应用于现有的医学图像检索方法中.

2 关联图模型

本节首先提出一种带有模糊性度量的关联图模型. 通过该模型,可以有效地对医学图像之间模糊的关联关系加以表示. 继而,给出基于关联图模型的医学图像 Top-*k* 查询问题的形式化定义.

2.1 数据模型

图模型通过将事物抽象为图中的点、事物之间的关系抽象为图中的边,可以有效地对事物及其之间的关联关系进行表示. 一般来说,一个(确定)有向图 *G* 定义为一个二元组 $G=(V,E)$,其中 *V* 是顶点集,*E* 是属于 $V\times V$ 的边集.

然而,在现实世界中事物之间的关联关系并不都可以进行准确的度量. 例如:已知图像 *A* 与图像 *B* 的相似性为 90%,图像 *B* 与图像 *C* 的相似性为 80%,图像 *A* 与图像 *C* 的相似性是多少? 显然,仅

通过以上 2 个条件不能够准确地给出图像 A 与图像 C 的相似性,但是经过推理和计算我们可以通过以上 2 个条件得到图像 A 与图像 C 的相似性范围,并且这一相似性范围在现实应用中具有很强的实际意义.但是,目前已有的确定图模型不能对这一模糊的关联关系进行表示,故本文基于一般(确定)有向图模型提出一种带有模糊性度量的关联图模型.

定义 1. 关联图 $G=(V,E,L,P)$ 是一个有向图.其中, V 是顶点集, E 是属于 $V \times V$ 的边集, $L: V \rightarrow \{1,2,3,\dots,n\}$ 为顶点分配标签函数, $P:E \rightarrow [0,1] \times [0,1]$ 是边的赋值函数.

在关联图 G 中, $e(i) \in E$ 的权值 $P(e(i)) = [P_{\min}^{(e(i))}, P_{\max}^{(e(i))}]$ 存在模糊性度量.图 2 给出了一个关联图的实例,在图 2 中关联图 $G=(V=\{V_1, V_2, V_3\}, E=\{\langle V_1, V_2 \rangle, \langle V_2, V_1 \rangle, \langle V_3, V_1 \rangle\}, L, P)$,其中 $P(\langle V_1, V_2 \rangle) = [0.8, 0.95]$, $P(\langle V_2, V_1 \rangle) = [0.9, 1]$, $P(\langle V_1, V_3 \rangle) = [0.4, 0.4]$.通过图 2 我们可以看出,点 V_1 与点 V_2 的关联关系存在模糊性度量,即不存在一个确定的值来衡量他们的关联关系,而点 V_1 与点 V_3 之间存在确定的关联关系.

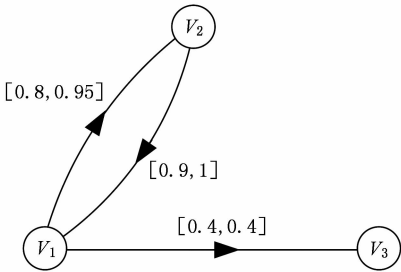


Fig. 2 Example of association graph.

图 2 关联图 G 示例

本文通过图像之间的相似性来定义图像之间的关联性.对于拥有 n 张图像的医学图像数据集 $M_{DB} = \{I_1, I_2, I_3, \dots, I_n\}$,根据关联图模型的定义,用于医学图像 Top- k 查询的关联图 G 的建模过程如下所示:

1) 对于医学图像数据集 $M_{DB} = \{I_1, I_2, I_3, \dots, I_n\}$ 中的每一张医学图像 I_i 在关联图 G 中创建点 V_i 与之相对应.

2) 在医学图像数据集 $M_{DB} = \{I_1, I_2, I_3, \dots, I_n\}$ 中,若存在图像 I_j 对于 I_i 的相似性度量,则在关联图 G 中增加边 $e(k) = \langle V_i, V_j \rangle$,且 $e(k) \in E$.其中边 $e(k) = \langle V_i, V_j \rangle$,是以 V_i 为起点、 V_j 为终点的有向边,表示以图像 I_i 为模板图像 I_j 对于图像 I_i 的相似性.

3) 在医学图像数据集 $M_{DB} = \{I_1, I_2, I_3, \dots, I_n\}$ 中,若存在图像 I_j 对于 I_i 的相似性为 a ,则对边 $\langle V_i, V_j \rangle$ 赋值 $P(\langle V_i, V_j \rangle) = [a, a]$;若图像 I_j 对于 I_i 的相似性介于 a 与 b 之间,其中 $a < b$,则对边 $\langle V_i, V_j \rangle$ 赋值 $P(\langle V_i, V_j \rangle) = [a, b]$.

可见,本文中关联图 G 的建模过程主要依赖于数据库中图像之间的相似性,这些相似性可以通过数据库中图像的历史查询结果获得.由于图像之间的相似性存在方向性(即以哪张图像为模板计算相似性),故本文在任何图像匹配过程中直接计算 2 张图像相互之间的相似性,从而减少不必要的重复匹配过程.为了便于描述,下文将使用点 V_i 代表图像 I_i .

由图 1 可以看出,医学图像在基于内容的检索过程中具有背景单一、图像内医学本体大小和位置相对固定等特点.基于以上特点,我们认为基于内容的医学图像匹配过程是一种基于图像内容特征的完全匹配,所以 $P(\langle V_i, V_j \rangle)$ 可形式化表示为

$$P(\langle V_i, V_j \rangle) = [P_{\min}^{(\langle V_i, V_j \rangle)}, P_{\max}^{(\langle V_i, V_j \rangle)}] = \left[\left(\frac{|V_i \cap V_j|}{|V_i|} \right)_{\min}, \left(\frac{|V_i \cap V_j|}{|V_i|} \right)_{\max} \right].$$

如在图 2 中, $P(\langle V_1, V_2 \rangle) = [0.8, 0.95]$,表示图像 V_2 对于图像 V_1 的相似性在 0.8 与 0.95 之间,即可形式化地表示为

$$P(\langle V_1, V_2 \rangle) = [P_{\min}^{(\langle V_1, V_2 \rangle)}, P_{\max}^{(\langle V_1, V_2 \rangle)}] = \left[\left(\frac{|V_1 \cap V_2|}{|V_1|} \right)_{\min}, \left(\frac{|V_1 \cap V_2|}{|V_1|} \right)_{\max} \right] = [0.8, 0.95].$$

同理 $P(\langle V_1, V_3 \rangle) = [0.4, 0.4]$ 可表示为 $\frac{|V_3 \cap V_1|}{|V_1|} = 0.4$.可见,关联图模型可以对医学图像之间确定以及模糊的相似性进行有效的表示.

2.2 问题定义

为了解决医学图像 Top- k 查询问题,最直接的方法是通过基于内容的医学图像检索方法比较查询图像和数据库中的所有图像,从而得到前 k 个最为相似的查询结果.但在现实世界中,用户往往只对前 k 个最为相似的查询结果感兴趣,可见,对医学图像数据库中图像逐张比较的 Top- k 查询方法将大部分时间浪费在与那些不可能具有高相似性图像的匹配过程中.本文旨在利用关联图模型,找到一种只需比较数据库中部分图像就可以得到与查询图像最为相似的前 k 个结果的 Top- k 查询方法.

关联图模型可以有效地对医学图像之间的相似性进行表示.本文首先给出基于关联图模型的医学图像 Top- k 查询问题的具体定义.

输入: 基于医学图像数据集 $M_{DB} = \{I_1, I_2, I_3, \dots, I_n\}$ 的关联图 $G = (V = \{V_1, V_2, \dots, V_n\}, E, L, P)$ 、查询图像 Q 、正整数 k 。

输出: 对于 $\forall i (1 \leq i \leq n)$, $P_{\min}^{(\langle Q, V_i \rangle)}$ 值最大的前 k 个结果。

根据关联图 G 的建模过程, 图 3 给出了一个在实际基于关联图模型进行医学图像检索过程中的关联图实例。在图 3 中, 点 V_1, V_2, V_3 及其之间的边是根据数据库中已有图像数据集 $M_{DB} = \{I_1, I_2, I_3\}$ 及其图像之间的相似性得到的。我们要做的即是最大程度地减少查询图像与数据库中已有图像的匹配次数。在图 3 中, 当有查询图像 Q , 选择图像 V_1 与其进行匹配, 得到 $P(\langle Q, V_1 \rangle) = [0.2, 0.2]$, $P(\langle V_1, Q \rangle) = [0.3, 0.3]$, 如何利用查询图像与数据库中图像的一次比较得到更多的有用信息, 从而尽快完成医学图像 Top- k 查询过程是本文研究的重点。

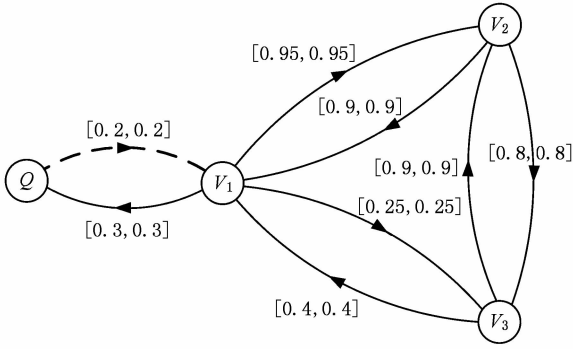


Fig. 3 Example of association graph uses in practical.

图 3 关联图 G 实例

3 关联性度量计算方法

本节将基于带有模糊性度量的关联图模型, 提出一系列关联性度量计算方法, 从而实现仅需对数据库中的图像比较一次便可获得查询图像与其他未匹配图像的更新。定理 1 给出了根据图像之间确定的相似性进行关联性度量的方法; 定理 2 给出了如何通过相似性范围对图像之间的关联性进行估计; 定理 3 给出了对关联图中边的相似性度量的更新规则。

定理 1. 对于关联图 G 中任意 2 个相邻接的点 V_i 和 V_j 以及查询图 Q , 如果有 $P(\langle Q, V_i \rangle) = [a, a]$, $P(\langle V_i, Q \rangle) = [b, b]$, 且 $P(\langle V_i, V_j \rangle) = [c, c]$, $P(\langle V_j, V_i \rangle) = [d, d]$, 则:

1) 边 $\langle Q, V_j \rangle$ 的最大值为

$$P_{\max}^{(\langle Q, V_j \rangle)}(a, b, c, d) = \begin{cases} 1, & b < c \text{ 且 } \frac{1-d}{1-b} \geq \frac{d}{c}; \\ a + (1-d) \times \frac{ac}{bd}, & b < c \text{ 且 } \frac{1-d}{1-b} < \frac{d}{c}; \\ 1 - a + \frac{ac}{b}, & b \geq c \text{ 且 } \frac{1-d}{1-b} \geq \frac{d}{c}; \\ \frac{ac}{bd}, & b \geq c \text{ 且 } \frac{1-d}{1-b} < \frac{d}{c}. \end{cases} \quad (1)$$

2) 边 $\langle Q, V_j \rangle$ 的最小值为

$$P_{\min}^{(\langle Q, V_j \rangle)}(a, b, c, d) = \begin{cases} \frac{ba - a + ac}{b}, & b > 1 - c; \\ 0, & b \leq 1 - c. \end{cases} \quad (2)$$

3) 边 $\langle V_j, Q \rangle$ 的最大值为

$$P_{\max}^{(\langle V_j, Q \rangle)}(a, b, c, d) = \begin{cases} 1, & b < c \text{ 且 } \frac{1-d}{1-b} \geq \frac{d}{c}; \\ 1 - d + \frac{bd}{c}, & b < c \text{ 且 } \frac{1-d}{1-b} < \frac{d}{c}; \\ d + \frac{(1-b)d}{c}, & b \geq c \text{ 且 } \frac{1-d}{1-b} \geq \frac{d}{c}; \\ 1, & b \geq c \text{ 且 } \frac{1-d}{1-b} < \frac{d}{c}. \end{cases} \quad (3)$$

4) 边 $\langle V_j, Q \rangle$ 的最小值为

$$P_{\min}^{(\langle V_j, Q \rangle)}(a, b, c, d) = \begin{cases} \frac{bd}{c} - \frac{d}{c} + d, & b > 1 - c; \\ 0, & b \leq 1 - c. \end{cases} \quad (4)$$

证明. 在关联图 G 中, 由 $P(\langle Q, V_i \rangle) = [a, a]$, $P(\langle V_i, Q \rangle) = [b, b]$, $P(\langle V_i, V_j \rangle) = [c, c]$, $P(\langle V_j, V_i \rangle) = [d, d]$ 可知:

$$\frac{|Q \cap V_i|}{|Q|} = a, \frac{|Q \cap V_i|}{|V_i|} = b, \\ \frac{|V_i \cap V_j|}{|V_i|} = c, \frac{|V_i \cap V_j|}{|V_j|} = d,$$

所以有 $a|Q| = b|V_i|$, $c|V_i| = d|V_j|$.

首先求 $P_{\max}^{(\langle Q, V_j \rangle)}$: 因为 $P(\langle Q, V_j \rangle)$ 可表示为 $\frac{|Q \cap V_j|}{|Q|} = \frac{|Q \cap (V_i \cap V_j)|}{|Q|} + \frac{|Q \cap (V_j - V_i)|}{|Q|}$, 又因为 $\frac{|Q \cap V_i|}{|Q|} = a$, 所以当 $\frac{|Q \cap V_i|}{|V_i|} < \frac{|V_i \cap V_j|}{|V_i|}$ 时, 即当 $Q \cap V_i$ 的点可以属于 $V_i \cap V_j$ 的真子集时, 有 $\frac{|Q \cap (V_i \cap V_j)|}{|Q|} \leq a$; 当 $\frac{|Q \cap V_i|}{|V_i|} \geq \frac{|V_i \cap V_j|}{|V_i|}$ 时, $\frac{|Q \cap (V_i \cap V_j)|}{|Q|} \leq \frac{|V_i \cap V_j|}{|Q|} = \frac{c|V_i|}{|Q|} = \frac{ac}{b}$. 又因为 $\frac{|V_i \cap V_j|}{|V_j|} = d$, 所以 $\frac{|V_j - V_i|}{|V_j|} = 1 - d$, 则有 $\frac{|V_j - V_i|}{|V_i|} =$

$(1-d) \times \frac{c}{d}$. 因为 $\frac{|Q \cap V_i|}{|V_i|} = b$, 所以 $\frac{|Q - V_i|}{|V_j|} = 1 - b$. 当 $\frac{|V_j - V_i|}{|V_i|} \geq \frac{|Q - V_i|}{|V_i|}$ 时, 即当 V_j 不相交于 V_i 的点多于 Q 不相交于 V_i 的点时, $\frac{|Q \cap (V_j - V_i)|}{|Q|} \leq \frac{|Q - V_i|}{|Q|} = 1 - a$; 当 $\frac{|V_j - V_i|}{|V_i|} < \frac{|Q - V_i|}{|V_i|}$ 时, $\frac{|Q \cap (V_j - V_i)|}{|Q|} \leq \frac{|V_j - V_i|}{|Q|} = \frac{|V_j - V_i|}{|V_i|} \times \frac{a}{b} = \frac{(1-d)ac}{bd}$. 综上所述, 得 $P^{(\langle Q, V_j \rangle)}$ 如式(1)所述.

同理可证得式(2)~(4), 由于篇幅限制, 本文不再赘述. 证毕.

例 1. 在图 4 中, 关联图 $G=(V=\{V_1, V_2, V_3\}, E, L, P)$. 当查询图像 Q 与 V_1 有 $P(\langle Q, V_1 \rangle)=[0.2, 0.2]$, $P(\langle V_1, Q \rangle)=[0.3, 0.3]$, 根据边 $P(\langle V_1, V_2 \rangle)=[0.95, 0.95]$, $P(\langle V_2, V_1 \rangle)=[0.9, 0.9]$, 通过定理 1 可得 $P(\langle Q, V_2 \rangle)=[0.1667, 0.2703]$, $P(\langle V_2, Q \rangle)=[0.2368, 0.3842]$. 同理, 通过边 $P(\langle V_1, V_3 \rangle)=[0.25, 0.25]$, $P(\langle V_3, V_1 \rangle)=[0.4, 0.4]$, 可得 $P(\langle Q, V_3 \rangle)=[0, 0.4167]$, $P(\langle V_3, Q \rangle)=[0, 1]$.

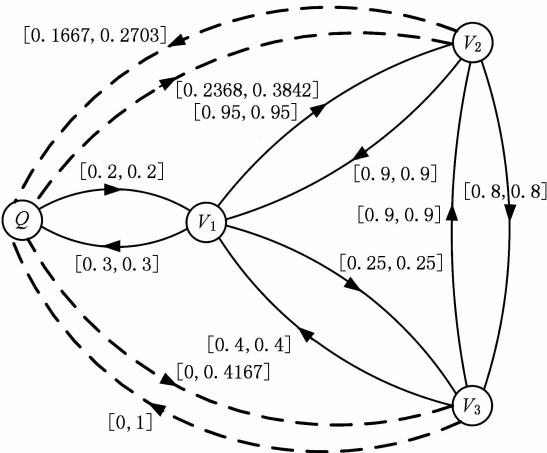


Fig. 4 The illustration of theorem 1.
图 4 定理 1 图例

通过例 1 我们可以发现, 查询图像 Q 通过与 V_1 比较一次便可得到 Q 与其他未查询图像的相似性范围, 即关联程度. 然而, 当查询图像 Q 与 V_1 相似性较低, V_1 与 V_2 相似性较低, 即当边 $\langle Q, V_1 \rangle, \langle V_1, Q \rangle, \langle V_1, V_3 \rangle, \langle V_3, V_1 \rangle$ 的得分较低时, 不能得到有效的 Q 与 V_3 的相似性范围, 如在例 1 中求得的 $P(\langle V_3, Q \rangle)=[0, 1]$.

定理 2. 在关联图 G 中, 若存在 $P(\langle V_i, V_j \rangle) =$

$[c, c], P(\langle V_j, V_i \rangle)=[d, d]$, 且查询图像 Q 与点 V_i 有 $P(\langle Q, V_i \rangle)=[P_{\min}^{(\langle Q, V_i \rangle)}, P_{\max}^{(\langle Q, V_i \rangle)}], P(\langle V_i, Q \rangle)=[P_{\min}^{(\langle V_i, Q \rangle)}, P_{\max}^{(\langle V_i, Q \rangle)}]$, 那么:

1) 边 $\langle Q, V_j \rangle$ 的最大值为 $P_{\max}^{(\langle Q, V_j \rangle)} = \max(P_{\max}^{(\langle Q, V_j \rangle)}(a_m, b_m, c, d))$, 其中 $a_m \in [P_{\min}^{(\langle Q, V_i \rangle)}, P_{\max}^{(\langle Q, V_i \rangle)}], b_m \in [P_{\min}^{(\langle V_i, Q \rangle)}, P_{\max}^{(\langle V_i, Q \rangle)}]$.

2) 边 $\langle Q, V_j \rangle$ 的最小值为 $P_{\min}^{(\langle Q, V_j \rangle)} = \min(P_{\min}^{(\langle Q, V_j \rangle)}(a_m, b_m, c, d))$, 其中 $a_m \in [P_{\min}^{(\langle Q, V_i \rangle)}, P_{\max}^{(\langle Q, V_i \rangle)}], b_m \in [P_{\min}^{(\langle V_i, Q \rangle)}, P_{\max}^{(\langle V_i, Q \rangle)}]$.

3) 边 $\langle V_j, Q \rangle$ 的最大值为 $P_{\max}^{(\langle V_j, Q \rangle)} = \max(P_{\max}^{(\langle V_j, Q \rangle)}(a_m, b_m, c, d))$, 其中 $a_m \in [P_{\min}^{(\langle Q, V_i \rangle)}, P_{\max}^{(\langle Q, V_i \rangle)}], b_m \in [P_{\min}^{(\langle V_i, Q \rangle)}, P_{\max}^{(\langle V_i, Q \rangle)}]$.

4) 边 $\langle V_j, Q \rangle$ 的最小值为 $P_{\min}^{(\langle V_j, Q \rangle)} = \min(P_{\min}^{(\langle V_j, Q \rangle)}(a_m, b_m, c, d))$, 其中 $a_m \in [P_{\min}^{(\langle Q, V_i \rangle)}, P_{\max}^{(\langle Q, V_i \rangle)}], b_m \in [P_{\min}^{(\langle V_i, Q \rangle)}, P_{\max}^{(\langle V_i, Q \rangle)}]$.

证明. 对于任意组 (a_m, b_m, c, d) , 其中 $a_m \in [P_{\min}^{(\langle Q, V_i \rangle)}, P_{\max}^{(\langle Q, V_i \rangle)}], b_m \in [P_{\min}^{(\langle V_i, Q \rangle)}, P_{\max}^{(\langle V_i, Q \rangle)}]$, 根据式(1)~(4)可求得 $P(\langle Q, V_j \rangle)=[u_m, v_m], P(\langle V_j, Q \rangle)=[x_l, y_l]$, 所以 $P(\langle Q, V_j \rangle)=[u_1, v_1] \cup [u_2, v_2] \cup \dots \cup [u_m, v_m], P(\langle V_j, Q \rangle)=[x_1, y_1] \cup [x_2, y_2] \cup \dots \cup [x_l, y_l]$, 所以 $P_{\max}^{(\langle Q, V_j \rangle)} = \max(u_1, u_2, \dots, u_l), P_{\min}^{(\langle Q, V_j \rangle)} = \min(v_1, v_2, \dots, v_l), P_{\max}^{(\langle V_j, Q \rangle)} = \max(x_1, x_2, \dots, x_l), P_{\min}^{(\langle V_j, Q \rangle)} = \min(y_1, y_2, \dots, y_l)$.

证毕.

例 2. 如图 5 所示. 在例 1 根据 $P(\langle Q, V_1 \rangle)=[0.2, 0.2], P(\langle V_1, Q \rangle)=[0.3, 0.3], P(\langle V_1, V_2 \rangle)=[0.95, 0.95], P(\langle V_2, V_1 \rangle)=[0.9, 0.9]$ 已计算出 $P(\langle Q, V_2 \rangle)=[0.1667, 0.2703], P(\langle V_2, Q \rangle)=[0.2368, 0.3842]$. 由定理 2, 根据 $P(\langle V_2, V_3 \rangle) =$

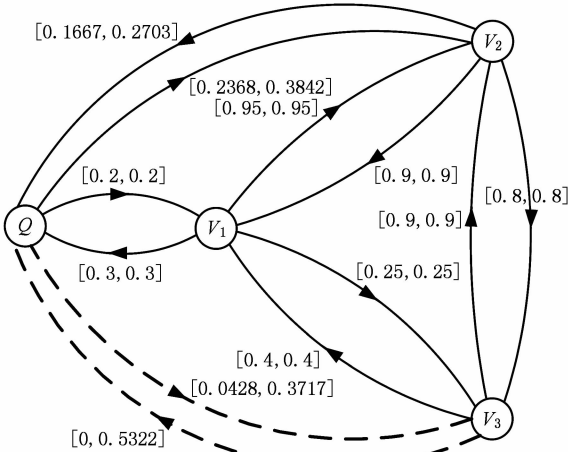


Fig. 5 The illustration of theorem 2.
图 5 定理 2 图例

$[0.8, 0.8], P(\langle V_3, V_2 \rangle) = [0.9, 0.9], P(\langle Q, V_2 \rangle) = [0.1667, 0.2703], P(\langle V_2, Q \rangle) = [0.2368, 0.3842]$ 可求得: $P(\langle Q, V_3 \rangle) = [0.0428, 0.3717], P(\langle V_3, Q \rangle) = [0, 0.5322]$. 该结果比例 1 中根据边 $\langle Q, V_1 \rangle, \langle V_1, Q \rangle, \langle V_1, V_3 \rangle, \langle V_3, V_1 \rangle$ 所求得的 $P(\langle Q, V_3 \rangle) = [0, 0.4167], P(\langle V_3, Q \rangle) = [0, 1]$ 更为精确.

通过例 2 我们发现, 定理 2 可以通过相似性范围以及关联图中更多的边来进行关联性度量, 可见定理 2 进一步扩大了关联性度量计算范围, 并且, 不同的推理路径对关联图 G 中同一条边的估计是不同的. 如在例 1 中, 通过点 $V_3 \rightarrow V_1 \rightarrow Q$ 对边 $\langle V_3, Q \rangle$ 的关联性度量是 $P(\langle V_3, Q \rangle) = [0, 1]$; 在例 2 中, 通过点 $V_3 \rightarrow V_2 \rightarrow Q$ 对边 $\langle V_3, Q \rangle$ 的关联性度量是 $P(\langle V_3, Q \rangle) = [0, 0.5322]$. 通过利用相似性较高的边, 关联性度量计算方法得到的推理结果更为精确. 可见, 定理 2 可以利用关联图中更多的边对某 2 张图像之间的关联性进行精确化.

定理 3. 在关联图 G 中若对边 $\langle V_i, V_j \rangle$ 存在不同的赋值 $P'(\langle V_i, V_j \rangle) = [P_{\min}^{(\langle V_i, V_j \rangle)}(a', b', c', d'), P_{\max}^{(\langle V_i, V_j \rangle)}(a', b', c', d')]$, $P''(\langle V_i, V_j \rangle) = [P_{\min}^{(\langle V_i, V_j \rangle)}(a'', b'', c'', d''), P_{\max}^{(\langle V_i, V_j \rangle)}(a'', b'', c'', d'')]$, 则有:

$$P(\langle V_i, V_j \rangle) = [\max(P_{\min}^{(\langle V_i, V_j \rangle)}(a', b', c', d'), P_{\min}^{(\langle V_i, V_j \rangle)}(a'', b'', c'', d'')), \min(P_{\max}^{(\langle V_i, V_j \rangle)}(a', b', c', d'), P_{\max}^{(\langle V_i, V_j \rangle)}(a'', b'', c'', d''))].$$

证明. 在关联图 G 中, 若存在 2 个对边 $\langle V_i, V_j \rangle$ 不同的赋值 $P'(\langle V_i, V_j \rangle) = [P_{\min}^{(\langle V_i, V_j \rangle)}(a', b', c', d'), P_{\max}^{(\langle V_i, V_j \rangle)}(a', b', c', d')]$, $P''(\langle V_i, V_j \rangle) = [P_{\min}^{(\langle V_i, V_j \rangle)}(a'', b'', c'', d''), P_{\max}^{(\langle V_i, V_j \rangle)}(a'', b'', c'', d'')]$. 根据赋值函数 P 的定义可知, $P(\langle V_i, V_j \rangle) = P'(\langle V_i, V_j \rangle) \cap P''(\langle V_i, V_j \rangle)$, 则边 $\langle V_i, V_j \rangle$ 的最小值

$$P_{\min}^{(\langle V_i, V_j \rangle)} = \max(P_{\min}^{(\langle V_i, V_j \rangle)}(a', b', c', d'), P_{\min}^{(\langle V_i, V_j \rangle)}(a'', b'', c'', d''));$$

根据定义, 边 $\langle V_i, V_j \rangle$ 的最大值

$$P_{\max}^{(\langle V_i, V_j \rangle)} = \min(P_{\max}^{(\langle V_i, V_j \rangle)}(a', b', c', d'), P_{\max}^{(\langle V_i, V_j \rangle)}(a'', b'', c'', d'')). \quad \text{证毕.}$$

例 3. 在例 1 中, 由定理 1 根据 $P(\langle Q, V_1 \rangle) = [0.2, 0.2], P(\langle V_1, Q \rangle) = [0.3, 0.3], P(\langle V_1, V_3 \rangle) = [0.25, 0.25], P(\langle V_3, V_1 \rangle) = [0.4, 0.4]$ 可求得对边 $\langle Q, V_3 \rangle$ 的赋值 $P(\langle Q, V_3 \rangle) = [0, 0.4167]$. 在例 2 中, 通过 $P(\langle V_2, V_3 \rangle) = [0.8, 0.8], P(\langle V_3, V_2 \rangle) = [0.9, 0.9], P(\langle Q, V_2 \rangle) = [0.1667, 0.2703], P(\langle V_2, Q \rangle) = [0.1842, 0.3842]$ 可求得对边 $\langle Q, V_3 \rangle$ 的另一赋值为 $P(\langle Q, V_3 \rangle) = [0.0428, 0.3717]$. 根

据定理 3, 可求得 $P(\langle Q, V_3 \rangle) = [\max(0, 0.0428), \min(0.4167, 0.3717)] = [0.0428, 0.3717]$. 同理可求得 $P(V_3, Q) = [\max(0, 0), \min(1, 0.5322)] = [0, 0.5322]$. 可见, 定理 3 定义了在关联图中根据不同的计算路径对图中的同一条边不同赋值的更新规则.

通过以上定理, 我们可以发现: 定理 1 给出了如何通过关联图中确定的相似性对 2 张图像之间的关联性进行度量; 定理 2 进一步给出了在关联图中利用相似性范围进行关联性度量的方法; 定理 3 给出了在关联图中对边赋值的更新规则. 通过该关联性度量计算方法, 当查询图像 Q 与 V_i 进行一次匹配, 即得到 $P(\langle Q, V_i \rangle)$ 与 $P(\langle V_i, Q \rangle)$, 则可通过关联图中的边对其他未匹配图像与查询图像 Q 的相似性进行推理. 该计算过程将有利于尽早地发现那些和查询图像 Q 不可能具有高相似度的图像, 从而减少图像检索过程中的图像匹配次数.

4 Top-k 查询方法及优化策略

本节根据关联性度量规则首先提出基于关联图模型的医学图像 Top-k 查询方法. 由于 Top-k 查询过程对匹配图像的选取敏感, 故本文根据 Top-k 查询方法的特点进一步提出 3 种基于游走的查询优化策略.

4.1 Top-k 查询方法

依据定理 1, 查询图像 Q 通过与 V_i 进行一次比较便可得到在关联图 G 中所有与 V_i 存在边 $\langle V_i, V_j \rangle, \langle V_j, V_i \rangle$ 的图像 V_j 与查询图像 Q 的相似性范围. 更进一步地, 定理 2 利用关联图 G 中更多的边对之前的推理结果进行精确化. 由此得到未匹配图像与查询图像 Q 的相似性范围, 从而挑选那些可能与查询图像 Q 具有高相似性的图像进行比较, 即对那些与查询图像 Q 不可能具有高相似性的图像进行剪枝.

由于本文所提出的关联图模型中, 边的表示采用模糊性度量方法, 即对边 $\langle V_i, V_j \rangle$ 的赋值为一个相似性范围 $P(\langle V_i, V_j \rangle) = [P_{\min}^{(\langle V_i, V_j \rangle)}, P_{\max}^{(\langle V_i, V_j \rangle)}]$. 也就是说, 对于关联图 G 中任意一点 V_i 存在 $P(\langle Q, V_i \rangle) = [P_{\min}^{(\langle Q, V_i \rangle)}, P_{\max}^{(\langle Q, V_i \rangle)}]$. 若图像 V_i 是与查询图像 Q 最为相似的第 k 个图像, 那么其他不属于 Top-k 结果的图像 V_j 对于查询图像 Q 的相似性最大值 $P_{\max}^{(\langle Q, V_j \rangle)} < P_{\min}^{(\langle Q, V_i \rangle)}$. 在本文中, 使用小顶堆 $minheap$ 来记录程序运行过程中与查询图像 Q 最为相似的

Top- k 结果,其中堆顶元素为 $minheap[1]$. 通过之前的假设,我们可以发现:1)当图像 V_j 与查询图像 Q 的 $P_{\min}^{(\langle Q, V_j \rangle)}$ 大于堆顶元素的最小值,图像 V_j 应该被插入 $minheap[]$ 中;2)当图像 V_j 与查询图像 Q 的 $P_{\max}^{(\langle Q, V_j \rangle)}$ 小于堆顶元素的最小值,应对图像 V_j 进行剪枝;3)当图像 V_j 与查询图像 Q 的 $P_{\max}^{(\langle Q, V_j \rangle)}$ 大于堆顶元素的最小值,图像 V_j 需进一步进行推理或计算.

由此,设计基于关联图模型的医学图像 Top- k 查询方法如算法 1 所示. 使用 $flag[i]$ 对图像 V_i 的状态进行表示, $flag[i]=1$ 表示图像 V_i 需要进一步判断, $flag[i]=0$ 表示图像 V_i 已被匹配或已被剪枝. 算法 1 的主要步骤为:1)查询图像 Q 随机选择 V_i , 与之比较得 $P(\langle Q, V_i \rangle)=[a, a], P(\langle V_i, Q \rangle)=[b, b]$;2)更新状态表 $qtov[i]=P(\langle Q, V_i \rangle), vtoq[i]=P(\langle V_i, Q \rangle)$, 若 $P_{\min}^{(\langle Q, V_i \rangle)}$ 大于堆顶元素 $minheap[1]$ 的最小值,更新堆;3)对于关联图 G 中每一张与 V_i 存在边 $\langle V_i, V_j \rangle, \langle V_j, V_i \rangle$ 的图像 V_j 执行迭代函数 $iteration(P(\langle Q, V_i \rangle), P(\langle V_i, Q \rangle), P(\langle V_i, V_j \rangle), P(\langle V_j, V_i \rangle))$;4)若仍存在图像 V_i 对于查询图像 Q 的最大值 $P_{\max}^{(\langle Q, V_i \rangle)}$ 大于堆顶元素 $minheap[1]$ 的最小值,执行步骤 1), 否则展示小顶堆 $minheap[]$ 里的结果并退出. 其中步骤 3)是一个迭代过程,如算法 2 所示.

算法 1. 基于关联图模型的医学图像 Top- k 查询算法(association graph reasoning, CGR).

输入: $G=(V=\{V_1, V_2, \dots, V_n\}, E, L, P)$ 、查询图像 Q 、正整数 k ;

输出:与查询图 Q 最为相似的前 k 个结果.

Int $qtov[n][2]=\{(0, 1)\}, gtoq[n][2]=\{(0, 1)\}, flag[n]=\{1\}, sum=1$;

初始化小顶堆 $minheap[k]$;

while ($sum>0$)

 随机选取图像 V_i , 如果 $flag[i]==1$;

$qtov[i]=P(\langle Q, V_i \rangle), vtoq[i]=P(\langle V_i, Q \rangle), flag[i]=0$;

 if ($qtov[i][1]>minheap[1]$)

 更新小顶堆;

 end if

 for ($j=1; j\leq n; j++$)

$iteration(P(\langle Q, V_i \rangle), P(\langle V_i, Q \rangle),$

$P(\langle V_i, V_j \rangle), P(\langle V_j, V_i \rangle))$;

 end for

 for ($i=1, sum=0; i\leq n; i++$)

$sum=sum+flag[i]$;

 end for

end while

输出小顶堆 $minheap[]$ 中保存的结果.

算法 2 给出了迭代函数 $iteration()$ 的主要步骤:1)由定理 1 或 2, 根据 $P(\langle Q, V_i \rangle), P(\langle V_i, Q \rangle), P(\langle V_i, V_j \rangle), P(\langle V_j, V_i \rangle)$ 计算 $P(\langle Q, V_j \rangle), P(\langle V_j, Q \rangle)$; 2)若 $P_{\min}^{(\langle Q, V_j \rangle)}$ 大于堆顶元素的最小值, 将图像 V_j 插入 $minheap[]$ 中; 3)若 $P(\langle Q, V_j \rangle), P(\langle V_j, Q \rangle)$ 与原状态不同, 更新状态表 $qtov[j]=P(\langle Q, V_j \rangle), vtoq[j]=P(\langle V_j, Q \rangle)$, 并对关联图 G 中每一张与 V_j 存在边 $\langle V_k, V_j \rangle, \langle V_j, V_k \rangle$ 的图像 V_k 执行迭代函数 $iteration(P(\langle Q, V_j \rangle), P(\langle V_j, Q \rangle), P(\langle V_k, V_j \rangle), P(\langle V_j, V_k \rangle))$, 否则退出. 通过该迭代过程, 可以根据关联性度量计算方法在关联图 G 中, 根据一次匹配结果尽可能多地对未匹配图像与查询图像 Q 之间的关联关系进行推测, 从而对那些与查询图像 Q 不可能具有高相似性的图像, 即那些不可能进入 Top- k 查询结果的图像进行剪枝.

算法 2. 迭代函数 $iteration()$.

输入: $P(\langle Q, V_i \rangle), P(\langle V_i, Q \rangle), P(\langle V_i, V_j \rangle), P(\langle V_j, V_i \rangle)$.

variation=0;

计算 $P(\langle Q, V_j \rangle), P(\langle V_j, Q \rangle)$;

if 需要更新 $qtov[j], vtoq[j]$

 更新 $qtov[j], vtoq[j]$;

 variation=1;

end if

if ($qtov[j][1]>minheap[1]$)

 更新小顶堆, $flag[j]=0$;

end if

if ($qtov[j][2]<minheap[1]$)

$flag[j]=0$;

end if

for ($k=1; k\leq n \& \& variation>0; j++$)

$iteration(P(\langle Q, V_j \rangle), P(\langle V_j, Q \rangle),$

$P(\langle V_j, V_k \rangle), P(\langle V_k, V_j \rangle))$;

end for

4.2 游走策略

如 4.1 节所述, 算法 1 只需对数据库中部分图像进行比较, 便可根据定理 1~3 得到查询图像 Q 与数据库中所有图像的相似性范围, 从而得到 Top- k 查询结果. 在算法 1 中, 采用随机选择策略对需要与查询图像 Q 进行匹配的图像 V_i 进行选择. 然而,

由于图像匹配的时间复杂度非常高,而选择不同图像与查询图像进行匹配将影响到运用定理 1~3 进行推理的过程.因此本节将对如何选取被匹配图像展开讨论,并提出了 3 种基于游走的查询优化策略:小顶堆游走策略、远离策略和基于聚类的游走策略.

4.2.1 小顶堆游走策略(CGR-heap)

在算法 1 中,小顶堆 $\text{minheap}[]$ 用以保存当前状态下与查询图像 Q 最为相似的前 k 个结果.由于本文所提出的 Top- k 查询方法基于关联性度量计算方法,所以在小顶堆中可能存在与查询图像 Q 未匹配的图像 V_i .在现实世界中,与查询图像 Q 具有低相似度的图像占图像总量的比例非常大,又根据定理 1~2,我们可以发现及早地与高相似度图像进行匹配,将有利于推理过程的精确化,从而减少需匹配图像次数.由此,提出小顶堆游走策略:

- 1) 如果小顶堆为空,随机选择图像 V_i 与查询图像 Q 进行匹配;
- 2) 如果小顶堆内存在图像未被匹配,选择与查询图像 Q 相似性较高的图像 V_i 与 Q 进行匹配;
- 3) 如果小顶堆内所有的图像均被匹配过,随机选择 V_i 与查询图像 Q 进行匹配.

4.2.2 远离策略(CGR-away)

在算法 1 中,当查询图像 Q 随机选择图 V_i 与其进行匹配后,将会根据关联性度量计算方法对查询图像 Q 和未与 Q 匹配的图像进行推理.由定理 2~3 可知,当查询图 Q 与 V_i 的相似性较低时,与 V_i 具有较高相似性的图 V_j 与查询图 Q 的相似性范围将迅速精确化.然而,若下次进行匹配的图像与 V_i 具有较高的相似性,则不能进一步扩大推理效果,增加比较需匹配图像的次数.因此,提出远离策略,其中心思想是远离前一张匹配图像.其主要步骤为:1)随机选择一张图像 V_i ,计算 $P(\langle Q, V_i \rangle)$, $P(\langle V_i, Q \rangle)$;2)挑选与前一张图像 V_i 相似度最低的图像进行匹配.

4.2.3 基于聚类的游走策略(CGR-cluster)

在信息检索等相关研究中,聚类方法往往被用于构建索引结构,从而快速找到查询信息所在的类,以减少比对次数.由定理 2~3 可知,查询图像 Q 与图像 V_i 进行一次匹配,可以迅速精确与图像 V_i 具有较高相似性的图像.我们也可以认为,查询图像 Q 与图像 V_i 进行一次匹配,可以迅速精确与图像 V_i 属于同一类的图像.因此,提出基于聚类的游走策略,并采用目前最为流行的 k -means 聚类方法进行聚类.

基于聚类的游走策略首先使用 k -means 方法将 M_{DB} 中的图像聚为 k 类,然后在执行算法 1 的过程中:

- 1) 从第 1 类到第 k 类,选择簇头 V_i 进行匹配;
- 2) 从第 1 类到第 k 类,在每个类中随机选择 1 张图进行匹配.

可以看出,基于聚类的游走策略令查询图像 Q 与数据库中图像的匹配以及推理过程较为均匀,从而确保 Top- k 查询方法的稳定性.

5 实验结果与分析

由于医学图像 Top- k 查询过程建立在有效的医学图像匹配的基础上,本文首先对比目前较为流行的 4 种基于内容的医学图像检索方法:LBP^[9], Sp-LBP^[8], KLT^[17], ULGR^[6].通过对这 4 种方法查准率的对比,选择查准率最高的方法从而实施本文提出的基于关联图模型的 Top- k 查询方法.为了进一步验证本文所提出 Top- k 查询方法的有效性,将其与相关的 Top- k 查询方法进行对比与分析.最后,对提出的基于游走的查询优化策略进行分析.

本文所使用的图像来自医院真实的脑部 CT 图像,分别取自 100 位病人,共 100×9 张真实图像,并在此基础上模拟了 4100 张图像^[18],共计 5000 张图像作为实验数据集.实验的硬件环境:Intel® Core™ 2 Quad CPU Q8400@2.66 GHz,2.0 GB RAM,采用 Microsoft Windows XP 操作系统;实验软件开发环境:VC 6.0.

5.1 现有基于内容医学图像检索技术对比

实验 1 分别实现了目前较为流行的 4 种基于内容的医学图像检索技术:LBP, Sp-LBP, KLT, ULGR.本文使用查准率来度量这 4 种基于内容的医学图像检索方法的准确性,其中:

$$\text{查准率} = \text{相似图像数} / \text{返回结果数}. \quad (5)$$

在本次实验中,我们假设人们只关心 Top-20 的返回结果,即返回结果数为 20.对于这 4 种基于内容的医学图像检索方法,本文分别进行了 50 次独立检索实验得到 4 种方法的查准率如图 6 所示.通过图 6,我们可以发现 ULGR 方法的平均查准率、查准率最大值、查准率最小值均高于其他 3 种方法.图 7 给出了这 4 种查询结果的实例,其中第 1 行图像为查询图像,其他 5 行为检索结果的 Top-5.在图 7 中,图像 67~70 是取自同一患者的脑部 CT 图像的相邻切层.通过实例可以发现,ULGR 方法的医学图像检索方法可以更为准确地找到与查询图像具有相似纹理的图像.

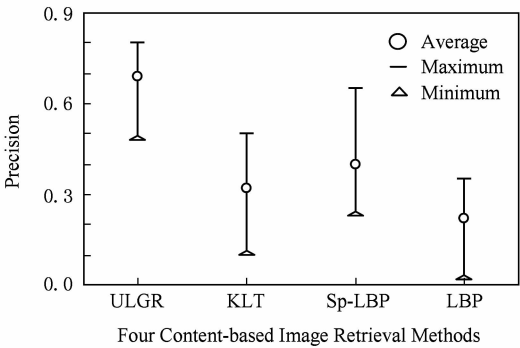


Fig. 6 Precision analysis for CBMIRs.
图6 医学图像检索技术查准率分析

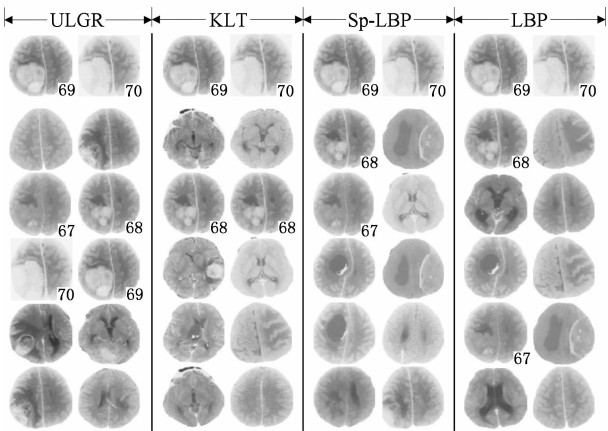
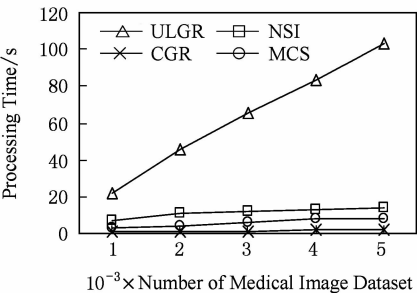
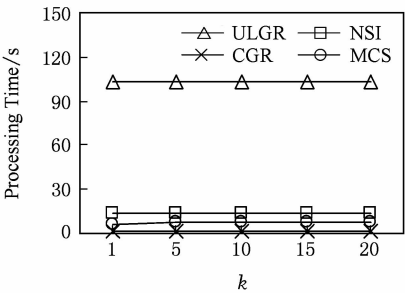


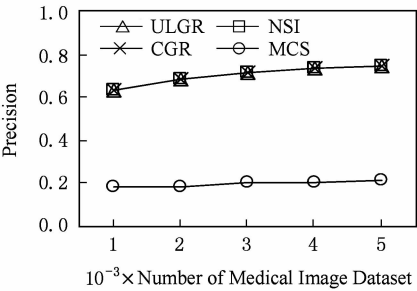
Fig. 7 Examples of searching results from CBMIRs.
图7 检索结果实例



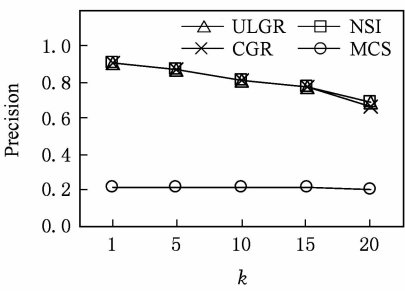
(a) Top-*k* methods processing time analysis 1



(b) Top-*k* methods processing time analysis 2



(c) Top-*k* methods precision analysis 1



(d) Top-*k* methods precision analysis 2

Fig. 8 Top-*k* methods performance analysis.
图8 Top-*k* 查询性能分析

5.2 Top-*k* 查询算法性能分析

由实验1可知,ULGR方法可以较为准确地对图像之间的相似性进行度量,故本文选取ULGR作为基本方法,从而实现Top-*k*查询过程.由于ULGR是基于不确定定点图的医学图像检索方法,故本实验选取与之相关的基于子图索引的Top-*k*查询方法(neighborhood subgraph index, NSI)^[19]、基于最大公共子图的Top-*k*查询方法(maximum common subgraph, MCS)^[20]与本文提出的CGR方法进行了对比.

图8(a)为对ULGR,NSI,CGR,MCS这4种方法分别经过50次独立实验测得的Top-20查询时间分析图.如图8(a)所示,这4种方法随图像数量的增加,查询时间均有所增加;但是,CGR方法时间消耗最低,且随着图像数量的增加变化最为平稳;ULGR方法耗时最高,且随图像数量的增加接近于线性增长.图8(b)对比随*k*值增加在5000张数据集上4种方法的Top-*k*查询时间的变化情况.其中ULGR,NSI方法随*k*值变化检索时间不变;MCS方法随*k*值增加查询时间有所增加;CGR方法随*k*值增加查询时间降低,且对于不同的*k*值其查询时间均低于其他3种方法.

在该实验中,同样使用查准率评判Top-*k*查询方法的准确性.如图8(c)所示,ULGR,NSI,CGR这3种方法查准率相同,且随图像数量的增加查准率

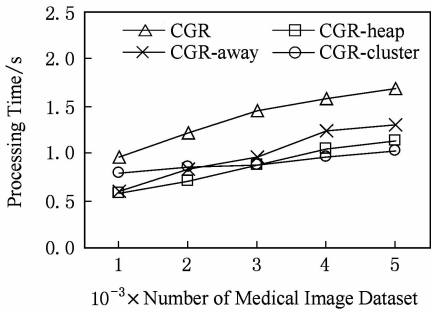
提高,因此,3 条线重合在一起;MCS 方法的查准率远低于其他方法,且随图像数量的增加查准率变化不大.图 8(d)为随 *k* 值变化 4 种方法的查准率变化情况,其中 ULGR,NSI,AGR 这 3 种方法查准率相同,且随 *k* 值增加查准率降低;MCS 方法的查准率远低于其他 3 种方法,且随 *k* 值增加查准率变化不大.综上所述,CGR 方法在整体性能上优于其他 3 种方法.

5.3 游走策略分析

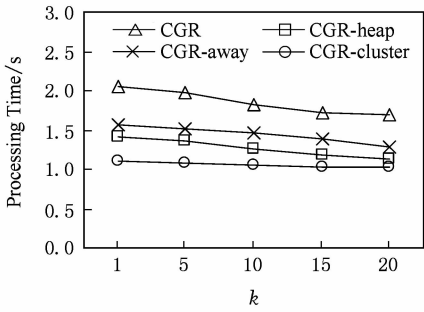
实验 3 对基于随机游走策略的 CGR 和 CGR-heap,CGR-away,CGR-cluster 方法进行比较.图 9(a)为经过 50 次独立实验的 4 种方法的查询时间分析图,其中 *k*=20.如图 9(a)所示,4 种方法的查询时间随图像数量的增长有所增加,其中 CGR 方法在不同规模数据集下的查询时间均大于其他 3 种方法.当数据集小于 3 000 时,CGR-heap 方法的查询时间最低;当数据集大于 3 000 时,CGR-cluster 方

法的查询时间最低.通过图 9(a)可以看出,CGR-cluster 方法随图像数量的增加查询时间增长最为平稳.图 9(b)为在 5 000 张数据集上 4 种方法取不同 *k* 值的查询时间比较图,可以看出 CGR-cluster 方法在不同 *k* 值的查询时间均低于其他 3 种方法,且随 *k* 值增加查询时间变化最为平稳.

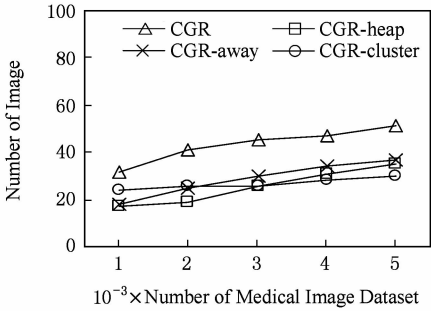
图 9(c)为 4 种方法的图像匹配次数分析图,其中 *k*=20.如图 9(c)所示,CGR 方法的图像匹配次数大于其他 3 种方法.在数据集小于 3 000 时,CGR-heap 方法的图像匹配次数最低;当数据集大于 3 000 时 CGR-cluster 方法的图像匹配次数最低.图 9(d)为在 5 000 张数据集上 4 种方法取不同 *k* 值的图像匹配次数比较图,可以看出 CGR-cluster 方法在不同 *k* 值的图像匹配次数均低于其他 3 种方法,且随 *k* 值增加图像匹配次数变化最为平稳.综上所述,CGR-cluster 对于大规模数据集在整体性能上优于其他 3 种方法.



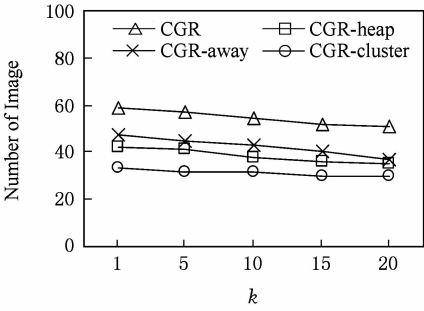
(a) Walk strategies processing time analysis 1



(b) Walk strategies processing time analysis 2



(c) Number of image comparison analysis 1



(d) Number of image comparison analysis 2

Fig. 9 Walk strategies performance analysis.

图 9 游走策略分析

6 总 结

本文研究了基于内容的医学图像检索方法的 Top-*k* 查询问题,首次使用带有模糊性度量的关联图模型对医学图像之间的关联关系进行表示;通过

提出一系列关联性度量计算方法,有效地减少了基于关联图模型的 Top-*k* 查询(CGR)过程中需要进行匹配的图像数量;最后,对如何选择被匹配图像的游走策略进行了讨论,并提出了小顶堆游走策略、远离策略和基于聚类的游走策略.实验结果表明,本文提出的 Top-*k* 查询算法,具有查询时间快、对图像

数量增长不敏感、准确率高等优势;并且通过对4种游走策略的比较,发现基于聚类的游走策略在平均搜索时间方面最短,并且对每次查询过程所需匹配图像数目较为稳定.

参 考 文 献

- [1] Deserno T M. Biomedical Image Processing [M]. Berlin: Springer, 2011: 471-494
- [2] Stephens S, Martin I, Dixon A K. Errors in abdominal computed tomography [J]. Journal of Medical Imaging, 1989, 3(5): 281-287
- [3] Akgül C B, Rubin D L, Napel S, et al. Content-based image retrieval in radiology: Current status and future directions [J]. Journal of Digital Imaging, 2011, 24(2): 208-222
- [4] Wiemker R, Dippel S, Stahl M, et al. Automated recognition of the collimation field in digital radiography images by maximization of the Laplace area integral [C] // Proc of Medical Imaging 2000: Image Processing. Bellingham, WA: SPIE, 2000: 1555-1565
- [5] Lehmann T M, Goudarzi S, Linnenbrügger N I, et al. Automatic localization and delineation of collimation fields in digital and film-based radiographs [C] //Proc of Medical Imaging 2002: Image Processing. Bellingham, WA: SPIE, 2002: 1215-1223
- [6] Pan Haiwei, Li Pengyuan, Li Qing, et al. Brain CT image similarity retrieval method based on uncertain location graph [J]. IEEE Journal of Biomedical and Health Informatics, 2014, 18(2): 574-584
- [7] Akgül C B, Rubin D L, Napel S, et al. Content-based image retrieval in radiology: Current status and future directions [J]. Journal of Digital Imaging, 2011, 24(2): 208-222
- [8] Unay D, Ekin A, Jasinschi R S. Local structure-based region-of-interest retrieval in brain MR images [J]. IEEE Trans on Information Technology in Biomedicine, 2010, 14(4): 897-903
- [9] Caetano T S, McAuley J J, Cheng L, et al. Learning graph matching [J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2009, 31(6): 1048-1058
- [10] Xu X, Lee D J, Antani S, et al. A spine X-ray image retrieval system using partial shape matching [J]. IEEE Trans on Information Technology in Biomedicine, 2008, 12(1): 100-108
- [11] Ilyas I F, Beskales G, Soliman M A. A survey of top- k query processing techniques in relational database systems [J]. ACM Computing Surveys, 2008, 40(4): Article No. 11
- [12] Wang Yijie, Li Xiaoyong, Qi Yafei, et al. Uncertain data queries technologies [J]. Journal of Computer Research and Development, 2012, 49(7): 1460-1466 (in Chinese)
- (王意洁, 李小勇, 祁亚斐, 等. 不确定数据查询研究[J]. 计算机研究与发展, 2012, 49(7): 1460-1466)
- [13] van Leuken R H, Garcia L, Olivares X, et al. Visual diversification of image search results [C] //Proc of the 18th Int Conf on World Wide Web. New York: ACM, 2009: 341-350
- [14] Chaudhuri S, Gravano L, Marian A. Optimizing top- k selection queries over multimedia repositories [J]. IEEE Trans on Knowledge and Data Engineering, 2004, 16(8): 992-1009
- [15] Lou Y, Snavely N, Gehrke J. MatchMiner: Efficient spanning structure mining in large image collections [C] // Proc of the 12th European Conf on Computer Vision. Berlin: Springer, 2012: 45-58
- [16] Arampatzis A, Zagoris K, Chatzichristofis S A. Dynamic two-stage image retrieval from large multimedia databases [J]. Information Processing & Management, 2013, 49(1): 274-285
- [17] Birchfield S. Derivation of Kanade-Lucas-Tomas tracking equation [OL]. [2014-03-01]. <http://www.ces.clemson.edu/~stb/klt>
- [18] Sled J G, Zijdenbos A P, Evans A C. A nonparametric method for automatic correction of intensity nonuniformity in MRI data [J]. IEEE Trans on Medical Imaging, 1998, 17(1): 87-97
- [19] Zhang Shuo, Gao Hong, Li Jianzhong, et al. Efficient query processing on uncertain graph databases [J]. Chinese Journal of Computers, 2009, 32(10): 2066-2079 (in Chinese)
(张硕, 高宏, 李建中, 等. 不确定图数据库中高效查询处理 [J]. 计算机学报, 2009, 32(10): 2066-2079)
- [20] Zhu Y, Qin L, Yu J X, et al. Finding top- k similar graphs in graph databases [C] //Proc of the 15th Int Conf on Extending Database Technology. New York: ACM, 2012: 456-467



Li Pengyuan, born in 1989. Received his MSc degree from Harbin Engineering University. His main research interests include medical data mining, computer-aided diagnosis.

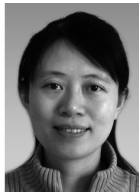


Pan Haiwei, born in 1974. Received his PhD degree from Harbin Institute of Technology. Senior member of China Computer Federation. Currently associate professor at Harbin Engineering University.

His research interests include database, data mining, massive data processing.



Li Qing, born in 1962. Received his PhD degree from the University of South California. Senior member of China Computer Federation. Currently professor at City University of Hong Kong. His research interests include object modeling, multimedia database and social media(itqli@cityu. edu. hk).



Xie Xiaoqin, born in 1973. Received her PhD degree from Tsinghua University. Member of China Computer Federation. Currently associate professor at Harbin Engineering University. Her research interests include data mining and social network analysis (xiexiaoqin@hrbeu. edu. cn).



Han Qilong, born in 1974. Received his PhD degree from Harbin Institute of Technology. Member of China Computer Federation. Currently associate professor at Harbin Engineering University. His research interests include spatiotemporal data mining and sensitive data protection(hanqilong@hrbeu. edu. cn).



Zhang Zhiqiang, born in 1973. Received his PhD degree from Tsinghua University. Senior member of China Computer Federation. Currently professor at Harbin Engineering University. His research interests include Web information retrieval, database technology(zqzhang@hrbeu. edu. cn).

2013 年《计算机研究与发展》高被引论文 TOP10

排名	论文信息
	孟小峰, 慈祥. 大数据管理:概念、技术与挑战[J]. 计算机研究与发展, 2013, 50(1): 146-169
1	Meng Xiaofeng and Ci Xiang . Big Data Management: Concepts, Techniques and Challenges [J]. Journal of Computer Research and Development, 2013, 50(1): 146-169
	傅颖勋, 罗圣美, 舒继武. 安全云存储系统与关键技术综述[J]. 计算机研究与发展, 2013, 50(1): 136-145
2	Fu Yingxun, Luo Shengmei, and Shu Jiwu. Survey of Secure Cloud Storage System and Key Technologies. Journal of Computer Research and Development, 2013, 50(1): 136-145
	李建中, 刘显敏. 大数据的一个重要方面:数据可用性[J]. 计算机研究与发展, 2013, 50(6): 1147-1162
3	Li Jianzhong and Liu Xianmin. An Important Aspect of Big Data: Data Usability [J]. Journal of Computer Research and Development, 2013, 50(6): 1147-1162
	刘大有, 陈慧灵, 齐红, 杨博. 时空数据挖掘研究进展[J]. , 2013, 50(2): 225-239
4	Liu Dayou, Chen Huiling, Qi Hong, and Yang Bo. Advances in Spatiotemporal Data Mining. Journal of Computer Research and Development, 2013, 50(2): 225-239
	廖彬, 于炯, 孙华, 年梅. 基于存储结构重配置的分布式存储系统节能算法[J]. 计算机研究与发展, 2013, 50(1): 3-18
5	Liao Bin, Yu Jiong, Sun Hua, and Nian Mei. Energy-Efficient Algorithms for Distributed Storage System Based on Data Storage Structure Reconfiguration [J]. Journal of Computer Research and Development, 2013, 50(1): 3-18
	贾冬艳, 张付志. 基于双重邻居选取策略的协同过滤推荐算法[J]. 计算机研究与发展, 2013, 50(5): 1076-1084
6	Jia Dongyan and Zhang Fuzhi. A Collaborative Filtering Recommendation Algorithm Based on Double Neighbor Choosing Strategy [J]. Journal of Computer Research and Development, 2013, 50(5): 1076-1084
	武建佳, 赵伟. WInternet: 从物网到物联网[J]. 计算机研究与发展, 2013, 50(6): 1127-1134
7	Wu Jianjia and Zhao Wei. WInternet: From Net of Things to Internet of Things [J]. Journal of Computer Research and Development, 2013, 50(6): 1127-1134
	余凯, 贾磊, 陈雨强, 徐伟. 深度学习的昨天、今天和明天[J]. 计算机研究与发展, 2013, 50(9): 1799-1804
8	Yu Kai, Jia Lei, Chen Yuqiang, and Xu Wei. Deep Learning: Yesterday, Today, and Tomorrow [J]. Journal of Computer Research and Development, 2013, 50(9): 1799-1804
	熊金波, 姚志强, 马建峰, 李凤华, 李琦. 基于行为的结构化文档多级访问控制[J]. 计算机研究与发展, 2013, 50(7): 1399-1408
9	Xiong Jinbo, Yao Zhiqiang, Ma Jianfeng, Li Fenghua, and Li Qi. Action-Based Multilevel Access Control for Structured Document [J]. Journal of Computer Research and Development, 2013, 50(7): 1399-1408
	张振海, 李士宁, 李志刚, 陈昊. 一类基于信息熵的多标签特征选择算法[J]. 计算机研究与发展, 2013, 50(6): 1177-1184
10	Zhang Zhenhai, Li Shining, Li Zhigang, and Chen Hao. Multi-Label Feature Selection Algorithm Based on Information Entropy [J]. Journal of Computer Research and Development, 2013, 50(6): 1177-1184

数据来源:中国知网;统计日期:2014-12-18