空间数据库中的关键词搜索

Ian De Felipe Vagelis Hristidis Naphtali Rishe

计算机与信息科学学院

佛罗里达国际大学

迈阿密，佛罗里达州 33199

{ian.de.felipe, vagelis, [rishen}@cis.fiu.edu](mailto:rishen%7d@cis.fiu.edu)

# 摘要：

许多应用需要寻找最接近包含有一系列关键词的位置的对象。比如，在线黄页允许用户指定一个地址和一系列关键词。然后，用户可以获取到符合条件的公司列表，它们按到用户指定地点的距离从近到远排序。空间数据的最近邻搜索问题和文本数据的关键词搜索问题已经被广泛地单独研究。然而，就我们所知，还没有一种方法能够有效地解决空间关键词搜索问题，即根据指定的位置和关键词来联合查询。

在这个研究中，我们提出了一种有效的方法，来解决top-k关键词查询问题。为此，我们引入了一个索引结构，IR2树（Information Retrieval R-Tree），它结合了R树和叠加文本签名。我们提出的算法，通过建立和维持一个IR2树，用来解决top-k空间关键词查询问题。通过实验比较，我们的算法与现有的方法相比，具有更优越的性能和良好的可扩展性。

# 1. 引言

越来越多的应用需要高效地执行被空间对象的属性限制的最近邻查询。理想的结果是，按照到查询区域的距离和描述文字与搜索关键词的相关性排序的对象。一个正在使用的简化的变体是距离优先的空间关键词查询，它将结果按距离排序，而关键词则作为筛选器来排除不包含它们的结果。

这里有一个运行中的例子，如Figure 1所示，给出了一些虚拟酒店的地理坐标和描述属性（名称、设施），一个空间关键词查询的示例是寻找离[30.5, 100.0]这个点最近，并包含关键词internet和pool的酒店。可以看到，最好的结果是酒店H7。

不幸的是，不能有效地支持top-k空间关键词查询，因为这时需要的是结果列表中的前几个。目前的系统采用ad-hoc结合最近邻和关键词搜索技术来解决这个问题。例如，R-Tree用来找到最近的邻节点，而对于每个邻节点，用倒排索引来检查是否包含查询的关键词。我们发现，这样分为两阶段处理的方法是低效的。我们提出了一个方法，能够有效地解决前k空间关键字查询问题，它是基于的紧密集成的数据结构、空间数据库搜索的算法和信息检索（IR）。需要特别说明的是，我们的方法要建立一个信息检索R-Tree（IR2-Tree），这是一个基于R-Tree的数据结构[Gut84]，查询时采用一种增量算法，使用IR2-Tree来高效地产生查询结果。

IR2-Tree是一种特殊的R-Tree，它的每个节点v都有一个签名（Faloutsos and

Christodoulakis [FC84]），用来表示以v为根的子树中所以空间物件的文本内容。我们的top-k空间关键词搜索算法 ，受到了Hjaltason和Samet[HS99]的工作的启发，运用他们的研究成果，我们通过访问很小的一部分IR2-Tree就可以找出前面的查询结果。我们的研究有如下成果：

* 定义了top-k空间关键词查询问题
* 提出了IR2-Tree——一个高效的索引结构，用来存储一系列对象的空间和文本信息。同时提出了IR2-Tree的维护算法，即插入和删除对象的算法。
* 提出了使用IR2-Tree的高效的增量算法，用来解决top-k空间关键词查询问题。我们用真实的数据进行试验，将它的性能与现有的方法相比较，在执行时间上有明显的提升。请注意，我们的方法可以用于任意形状和多维对象，为了使结果更加清晰，我们在实验中不仅使用了这两种维度上的点。

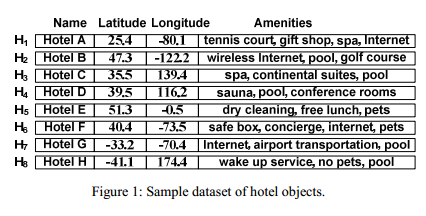
这篇论文按如下结构组织：第2节正式定义了tok-k空间关键词搜索问题；第3节介绍背景知识；第4节介绍IR2-Tree及其维护算法；第5节介绍增量搜索算法和其他基线算法；第6节通过实验比较我们的算法和其他基线算法；第7节讨论相关研究；第8节总结。

# 2. 问题定义

在这里，（空间）对象T被定义为有序对（T.p，T.t），T.p是多维空间中的位置描述符，T.t是一个文本文档（文本描述）。设D是数据库中所有对象的集合。在Figure 1中，T.p是由“纬度（latitude）”和“经度（longitude）”组成的点，而T.t是“旅馆名（name）”和“设施（amenities）”属性的级联。

top-k空间查询Qs，是指搜索多维空间，来找到离指定点p最近的k个对象。这些空间对象按距离排序，使得离p越近的点排位越靠前。特别地，

score(T) = distance(T.p, p)。例如，在Figure 1中，对于给定的p[30.5, 100.0]，对象H4排在最前面。



关键词查询Qw，是一系列关键词的集合w1,…wm。Qw的返回值对象的列表，按它们的文本描述和关键词的相关度排序，其中相关度IRscore(T.t, Qw)通过一个IR排序函数[Sin01]计算。

在运行实验中使用的一个特殊的例子，是一个布尔关键词查询，返回文本文档中包含有全部关键词w1,…wm的对象，即：例如，在Figure1中，对象H2和H7是布尔关键词查询{“internet”, “pool”}的结果。

Top-k空间关键词查询Q是top-k空间查询和关键词查询的结合，它被定义为需要的结果数Q.k，指定点Q.p，一个关键词的集合Q.t={w1,…,wm}和一个排序函数f(distance(T.p, Q.p), IRscore(T.t, Q.t))。查询Q的结果是一个top-k对象的列表T，这k个对象根据排序函数f排序。

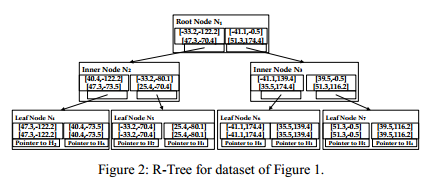
一种特殊的情况是距离优先tok-k空间关键词查询Q，在我们的实验中也使用了，它返回包含w1,…,wm且最接近Q.p的k个对象。也就是说，距离优先tok-k空间关键词查询是top-k空间关键词查询和布尔关键词查询的组合。即：



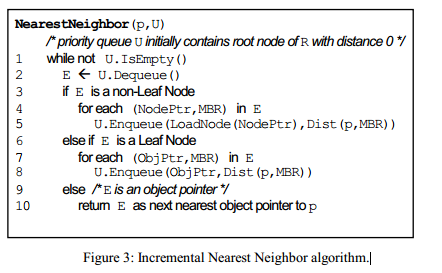
比如，在Figure1中，对象H7和H2是Q.k=2，Q.p=[30.5,100]，和Q.t={“internet”, “pool”}距离优先tok-k空间关键词查询的结果。我们的工作解决了高效解答top-k空间关键词查询的问题。

# 3. 增量最近邻法的背景知识

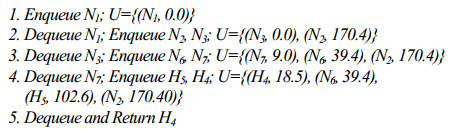
Figure2所示的是R-Tree的一个例子，使用的是Figure1中的旅馆数据。MBR用西南和东北的点表示。R-Tree通常存储在磁盘上，每个R-Tree占用一整个磁盘块，因此，访问节点需要进行磁盘I/O。节点所能指向的子节点数量，称为节点的容量。



Hjaltason and Samet [HS99]提出的增量最近邻算法，使用R-Tree连接最少的R-Tree节点和对象，并以增量的方式来获取离给定点或区域最近的对象。Figure3所示的是二维对象的增量最近邻算法。输入参数是查询点p（或者查询区域），和用R-Tree的根节点R初始化的优先队列U。第二行返回队列中离查询点距离最小的元素。如果这个元素是叶子节点，它的每个子对象通过ObjPtr引用，按距离插入到队列中；如果这个元素不是叶子节点，则它的每个子节点通过NodePtr引用，插入到队列中。最后，如果这个元素是空间对象的指针，则作为算法的下一个结果返回，如第10行所示。Dist函数计算查询点p和MBR之间的距离。我们假设R-Tree保存在磁盘中，因此，LoadNode函数从磁盘加载节点。



例1：在Figure2的R-Tree上执行增量最近邻算法，查询点是[30.5, 100.0]，执行的步骤如下：

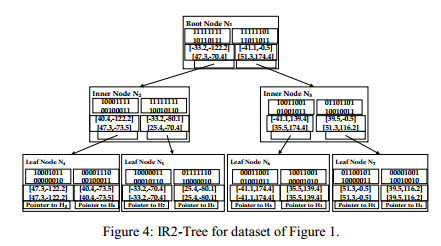


如果继续执行，接下来返回的结果就是H3, H5, H8, H6, H1, H7, H2。

# 4. IR2-Tree

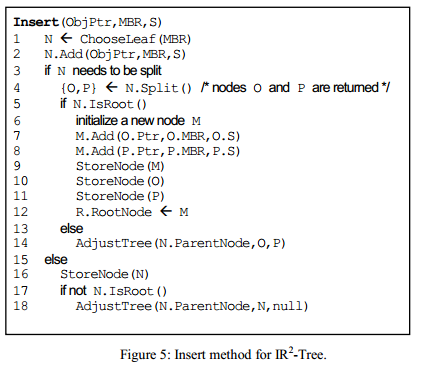
IR2-Tree是R-Tree和签名文档的结合。特别的是，IR2-Tree的每个节点都包含空间和关键词信息；前者在表中的最小边界区，后者在表中的签名中。IR2-Tree对top-k空间查询和top-k空间关键词查询都有利，这一点将在稍后说明。

更正式的说，IR2-Tree R是一种高度平衡树数据结构，其中每个叶子节点具有如下形式的条目(ObjPtr, A, S)，ObjPtr和A与R-Tree中定义的一样，S是ObjPtr引用的对象的签名。非叶子节点具有如下形式的条目(NodePtr, A, S)，NodePtr和A与R-Tree中定义的一样，而S是节点的签名。一个节点的签名是其条目的所有签名的叠加(OR-ing)。因此，一个节点的签名等效于在其子树中的所有文件的签名。Figure4所示的是Figure1中的简单数据集的IR2-Tree。为了简化下面的介绍，我们着重于二维空间。



IR2-Tree通过插入和删除操作来维护，它们都是在对应的R-Tree操作基础上修改的。Figure5和Figure6分别展示了IR2-Tree的插入和删除算法。

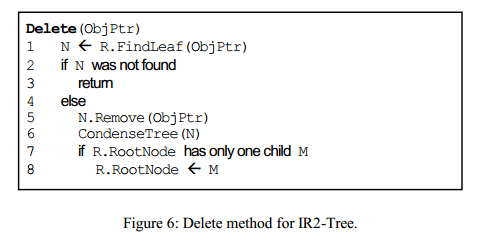
插入算法使用了一个标准的R-Tree方法ChooseLeaf，在[Gut84]中能够找到。使用标准Quadratic Split技术[Gut84]来做节点拆分。我们修改了标准的AdjustTree方法，使他能维护修改后的包含签名的节点。即，节点N的一个新的位被设为1时，它的祖先也必须设为1。最后，假定所有的树相关的算法，能够隐式地访问IR2-Tree R的根节点。



Insert算法的输入是对象T的指针，MBR和签名。算法的第一行根据T的MBR获取一个最合适的叶子节点。然后T的指针，MBR和签名存储到N中。如果N已经达到了它的最大容易，它将会被拆分。如第四行所写，如果N被拆分为节点O和P，且N为跟节点，则将创建一个新节点M，M称为O和P的父节点，保存它们的指针、MBR和签名。最后，M被声明为新的根节点。如果N不是跟节点，则它的父节点也要更新，像14行和18行写的那样。最后，由于我们假设IR2-Tree保存在磁盘中，StoreNode函数将节点保存到相应的磁盘块。

Delete的实现中使用了FindLeaf的标准实现，而CondenseTree则经过修改以维护修改的节点中的签名，与上面的AdjustTree类似。在Figure6的第一行，执行对包含有要删除的对象T的节点N的搜索，如果存在这样的节点N，则从节点N中移除对象T，否则算法结束。如果对象T被移除，树需要重新调整。

很明显，Insert和Delete算法的复杂度和R-Tree中是一样的，因为唯一的额外操作是维护修改的节点及其祖先节点的签名。需要注意的是，整个节点及其祖先节点的签名的更新与树更新节点及其祖先的MBR同时进行。为了弥补存储IR2-Tree的节点中的签名需要的额外的空间，同时为了与R-Tree拥有相同数量的子节点，我们在需要时为IR2-Tree的节点增加磁盘块。正如6.1节所示，这个事实对IR2-Tree算法的性能有轻微的影响。



多级IR2-Tree

上面描述的IR2-Tree有一个缺点，每一级都使用相同的签名长度，这会在高级别上带来更多的误报（因为它们是低级别的叠加）。为了解决这个问题，我们队不同的级别使用不同的签名长度。这通过使用多级叠加编码来实现，以减少误报，特别是在非叶子节点上。在这种情况下，每个节点都使用最佳签名长度（使用最佳签名长度公式计算[MC94]），并用节点的子树中所有对象的签名来叠加，而不是之前的子节点的签名。这种变化后的方法，称为多级IR2-Tree（MIR2-Tree）。它也有一个缺点，即显著地增加了维护操作（插入、删除）的复杂度。插入、删除每个对象，都必须访问相关的所有对象，来计算所有祖先节点的签名，而不是像之前那样，只需要计算子节点的签名。我们将在第6节中比较IR2-Tree和MIR2-Tree的性能。

# 5. 解决空间关键词查询问题的算法

在5.1节中，考虑两个基线算法，根据使用的底层数据结构分别命名为：R-Tree和唯一倒排索引（IIO）。在5.2节中，我们介绍距离优先IR2算法，它使用IR2-Tree数据结构来解决距离优先空间关键词查询问题。在5.3节中，我们介绍一般的IR2算法，它使用IR2-Tree数据结构来解决一般的空间关键词查询问题。值得注意的是，最后两种算法还能够在没有修改操作的MIR2-Tree使用。

## 5.1目前的基线算法

为了简单起见，我们只介绍针对简化的距离优先空间关键词查询问题的R-Tree和IIO基线算法，在第6节的实验中也将使用。这两种算法都可以通过扩展来解决一般的空间关键词查询问题。

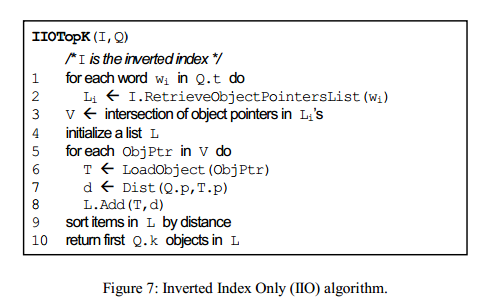
R-Tree算法

第一个基线算法，R-Tree，只使用了R-Tree数据结构。对于给定的距离优先空间关键词查询问题，这个算法首先找出离查询点Q.p最近的对象。然后反向检索该对象（由于R-Tree只包含对象指针），比较该对象的文本描述和查询关键词。如果匹配失败，则这个对象被丢弃，继续检索下一个最近的对象。这个过程使用了Figure 3中的增量最近邻算法[HS99]。程序持续运行，直到找到文本内容包含有查询关键词的对象。一旦找到符号条件的对象，就返回该对象，这个过程重复执行，直到返回k个对象。

这个算法的缺点是，必须检索最近邻算法返回的每个对象，直到找到top-k结果。这可能会导致对很多无用对象的检索。最坏的情况下（当没有一个对象匹配查询关键词时），整棵树都会被遍历，每个对象都要被检查一次。

IIO算法

IIO基线算法使用倒排索引。首先通过对倒排索引返回的列表取交集，来找到所有文本文档中包含查询关键词的对象（通过对象id）。设V为交集中对象的集合。然后检索V中的对象，计算每个对象与查询点Q.p之间的距离。随后，将这些对象排序，返回其中的top-k对象。IIO算法如Figure 7所示。输入的参数是倒排索引表I和距离有限top-k空间关键词查询Q。



例2：考虑如下查询，在Figure 1的数据上，进行包含有关键词{“internet”, “pool”}，查询点为[30.5, 100.0]的top-2旅馆查询。IIO算法的执行过程如下：

1. 关键词“internet”的倒排索引返回的结果是H1，H2，H6，H7
2. 关键词“pool”的倒排索引范湖的结果是H2，H3，H4，H7 ，H8
3. 取交集，结果是H2，H7
4. 访问H2，H7，获取坐标
5. 添加H2到列表L={(H2, 222.8)}
6. 添加H7到列表L={(H 7, 181.9), (H 2, 222.8)}，返回H2，H7

当有很多对象都包含查询关键词时，这个算法的性能会降低很多。在这种情况下，倒排索引会返回很多对象，需要一一检索和比较。值得注意的是，IIO算法是本文提到的算法中唯一一个不是增量的算法。也即是说，IIO算法计算所有的查询结果，正如第6节将要提到的那样，它的性能取决于k。

## **5.2 距离优先**IR2-Tree算法

在本节中，我们讨论距离优先版本的IR2-Tree算法，算法输出包含所有关键词的对象，并按它们到查询点的距离排序。在5.3节中，我们将展示如何扩展这个算法来解决一般的top-k空间关键词查询问题。

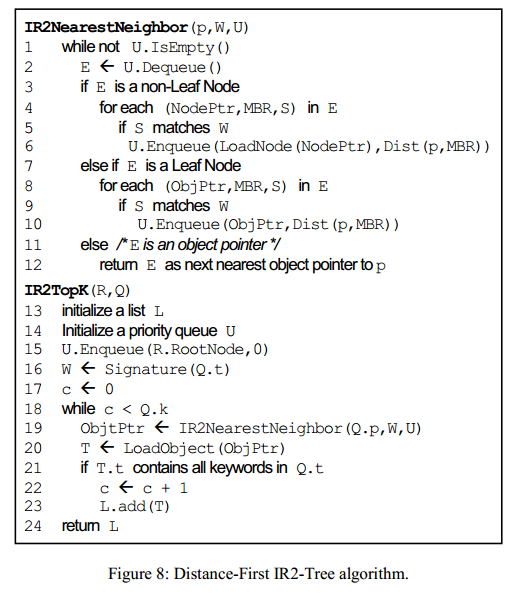
距离优先IR2-Tree算法利用IR2-Tree结构来高效地解决top-k空间关键词查询问题。树的遍历基于增量最近邻算法（Figure 3）。该算法的主要优点是：如果根节点的签名与查询签名Signature(Q.t)不匹配，将剪去整个子树。这样做的合理性在于，IR2-Tree的节点的签名是由其所有子节点的签名组成的。这个是传统增量最近邻法提供的修剪之外的修剪。通过这两种修剪机制的紧密结合，距离优先IR2-Tree算法能够访问尽量少的IR2-Tree节点和对象，来解答距离优先top-k空间关键词查询问题。

Figure 8显示了距离优先IR2-Tree算法(IR2TopK)。其中，最关键的方法是IR2NearestNeighbor(.)，它基于NearestNeighbor算法，在其上增加了一个表示查询的签名的输入变量w。节点和对象的签名将与w进行比较，如果不匹配将被跳过（即从搜索队列中移除）。需要注意的是，对IR2NearestNeighbor(.)方法的每次访问都返回一个候选的结果，它将在稍后（21行）进行检查，确保不会误报。

例3：在这个例子中，我们在Figure 4中的IR2-Tree上执行上述算法，来解决如下查询：top2旅馆查询，查询点[30.5, 100.0]，关键词{“internet”, “pool”}。

1. N1入队列，U={(N 1, 0.0)}；
2. N1出队列，N2入队列，U={(N 2, 170.4)}；
3. N2出队列，N4、N5入队列，U={(N 5, 170.5), (N 4, 173.8)}；
4. N5出队列，H7入队列，U={(N 4, 173.8), (H 7, 181.9)}；
5. N4出队列，H2入队列，U={(H 7, 181.9), (H 2, 222.8)}；
6. H7出队列并返回；
7. H2出队列并返回。

请注意，IR2-Tree的签名修剪能力如何开始在第2行出现。只有一个子节点N1入队列，其他的由于签名不匹配而被丢弃了。当对象H1 和H6的父节点被访问时，它们也被修剪了（在第4行和第5行）。



## **5.3** IR2-Tree算法

在本节中，我们讨论一般的IR2-Tree算法，对象通过第2节中定义的排序函数f(distance(T.p, Q.p), IRscore(T.t, Q.t))来排序输出。它与距离优先版本的算法的主要区别如下：

（i）不创建单独的签名Signature(Q.t)，而使用独特的签名Signature(w), w∈Q.t。原因是，这里不使用AND语义，即，只包含有查询关键词中的一部分的对象也可能出现在结果中。

（ii）不能再因为一个对象是下一个最近的，且包含所有关键词而立即输出它，因为，更远的对象也有可能有更高的整体分f(.)。因此，节点v在队列U中按它们所包含的对象的最大分值排序，即，



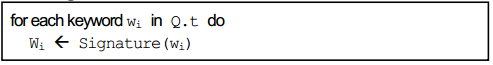
假设f()随着distance()增大而减小，随着IRscore()增大而增大，则有：



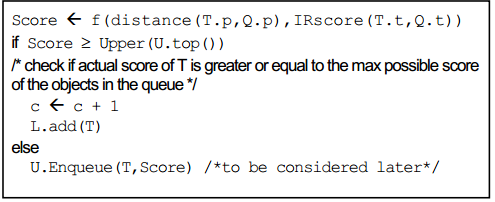
为了计算v的MBR中的对象的最大可能的IR分值UpperBoundT-has-signature-v.S(IRscore(T.t, Q.t))，假设v中有一个假想的对象T包含Q在v.S的签名中指定的所有关键词一次（词频tf=1），即，假设没有误报。因此，T.t的文档长度（dl）就是这些关键词的数量。接着可以使用tfidf IR排序函数[Sin01]。这种方法有利于尽早地输出结果对象。注意，如果IR函数使用了先进的功能，如词库和本体论等，则不可能估计一个很准确的最大可能IR分值。

我们队Figure 8中的距离优先版本做了如下修改：

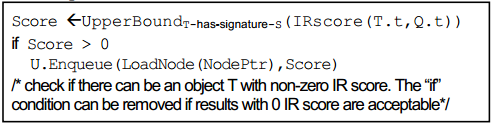
1. 替换16行:



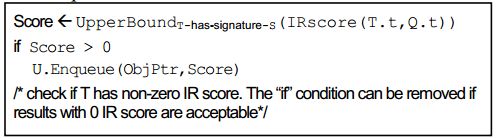
2. 替换21到23行:



3. 替换5到6行:



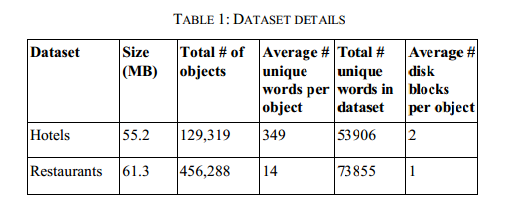
4.替换9到10行:



# 6. 实验

为了衡量IR2-Tree、MIR2-Tree和其它基线算法的效率，我们用JAVA实现了所有的算法和底层数据结构。其中，所有的索引结构（R-Tree、IR2-Tree、MIR2-Tree和倒排索引）都保存在磁盘中。实验主要针对距离优先版本的top-k空间关键词查询问题，因为它的结果比较容易理解和分析。空间对象保存在一个纯文本文件中，树的叶子节点中保存指向对象在文件中的位置的指针。我们根据满足查询时间和执行时间锁需要的磁盘访问来做比较。实验所用的机器配置是：CPU：Athlon 64 3400+ (NewCastle)，内存：2GB，硬盘驱动器：74GB 10000RPM。

我们使用的两个数据集均来自高性能数据研究中心（High Performance Database Research Center ，http://hpdrc.fiu.edu/）。两个数据集都是纯文本文件（制表符分割），每个空间对象占一行。第一个数据集包含的对象是一些有代表性的旅馆，作为旅馆数据集。第二个数据集作为餐馆数据集，包含餐馆数据。Table 1展示了这两个数据集的详细信息。



在所有的实验中，磁盘块大小是4096KB。同时，R-Tree的一个节点的子节点数量，根据每个节点占一个磁盘块计算。也就是说，在实现时，每个节点可以容纳113个子节点。对于IR2-Tree和MIR2-Tree，也使用的相同数量，因此，每个节点需要占两个磁盘块。实验表明，虽然IR2-Tree和MIR2-Tree需要额外的磁盘块开销，但是，这对执行时间的影响很小。

为了比较IR2-Tree、MIR2-Tree、R-Tree和IIO算法的性能，我们进行了三组实验。第一组实验测试当需要的结果数k不同时的算法性能；第二组实验测试查询关键词的数量的影响；最后，第三组实验测试签名长度r的影响。

k（top-k）的取值

在这组实验中，我们把查询关键词的数量固定位2，签名长度固定位189bytes（对于旅馆数据集）和8bytes（对于餐馆数据集）。需要注意的是，MIR2-Tree的顶层的签名长度会更长，它使用可变签名长度。这些签名长度的选取，需要使空间的需求和性能保持平衡。对旅馆和餐馆的实验结果分别在Figure 9和Figure 12中。图中Y轴使用对数标度，以使差异更直观。

从图中可以看出，IR2-Tree和MIR2-Tree在k的所有取值下，性能都比R-Tree好。这是因为R-Tree方法需要访问更多的对象和潜在的树节点。相反，IR2-Tree和MIR2-Tree使用签名来修剪整个子树。特别的，MIR2-Tree在过滤内部节点上做得更好，因为它对树的每一层采用了最佳签名长度，正如第4节所描述的。

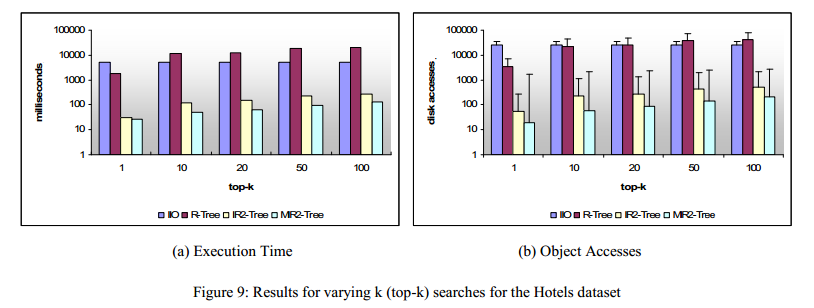
Figure 9(b)和12(b)显示了各个算法访问的磁盘块数量。柱状图表示随机磁盘块访问次数，而上面的细线表示顺序磁盘块访问次数。和预期的一样，执行时间主要和随机访问次数成正比。值得注意的是，由于修剪的效果，MIR2-Tree比IR2-Tree进行更少的随机磁盘访问，但顺序磁盘访问更多。这主要是因为MIR2-Tree的顶部节点由于签名更长而占用更多的磁盘块。IIO算法对k的取值不敏感，因为它需要检查包含所有关键词的所有对象。

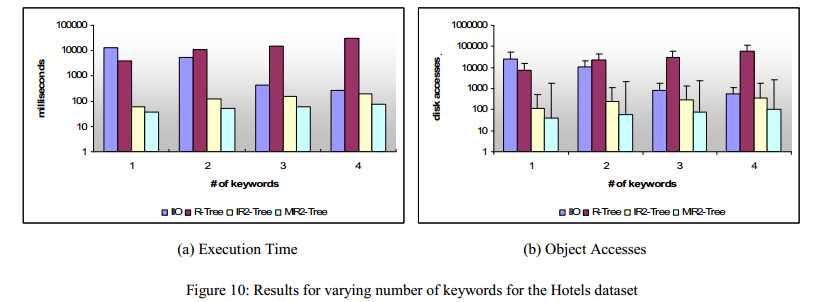
关键词数量

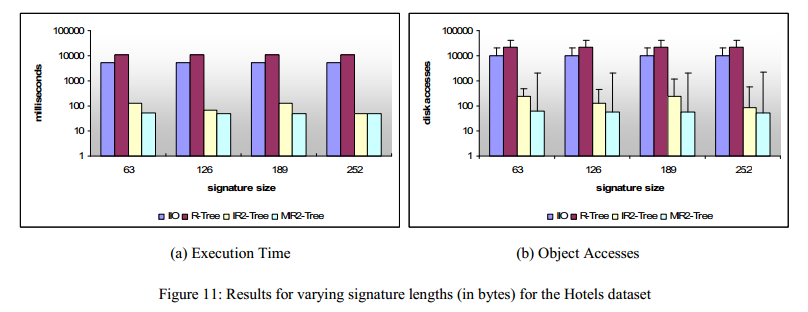
在这组实验中，我们把需要的对象数量k固定为10，签名长度跟上一组实验相同。实验结果请参阅Figure 10和Figure 13。在关键词数量增加的同时，包含全部关键词的对象也减少了（因为距离优先空间关键词查询时合取）。我们注意到，IIO算法的性能随着关键词数量增多而变好，这是由于倒排索引表的交集变小，需要访问的对象变少。

签名长度

在这组实验中，我们把k固定为10，关键词数量固定为2。实验结果请参阅Figure 11和Figure14。首先要注意，旅馆数据集选择的签名和餐馆数据集是不同的，这是因为一个旅馆对象包含更多的独特关键词，如Table 1所示。需要注意，在MIR2-Tree中，签名长度是指叶子节点使用的签名长度，而顶部节点则使用更长的签名。在增加IR2-Tree和MIR2-Tree的签名长度时，有一个权衡。增加签名长度，能够减少误判，但是也会增加树结构占用的空间，从而增加磁盘访问。因此，签名长度变化对实验结果的影响并不明确。







A. 需要的空间

Table 2显示了每个结构需要的总空间（以MB为单位）。对于IR2-Tree和MIR2-Tree，我们考虑的实例是前面“top-k的取值变化”和“关键词数量变化”时的实验，即，分别为189和8bytes。

餐馆数据集的IIO结构大小显著地小于旅馆数据集。这是因为，餐馆数据集中平均每个对象包含的独特关键词更少，如Table 1所示。另一方面，餐馆数据集的树结构比旅馆数据集更大，因为有更多的对象。

B. 讨论