Fatores Determinantes de Desempenho de Métodos de Acesso Multidimensionais

RICARDO RODRIGUES CIFERRI[1]
ANA CAROLINA SALGADO[2]
[1]DIN/UEM – Departamento de Informática da Universidade Estadual de Maringá,
Av. Colombo, 5790, Zona 07, 87020-900 Maringá, PR, Brasil
rrc@din.uem.br
[2]CIn/UFPE – Centro de Informática da Universidade Federal de Pernambuco,
Caixa Postal 7851, Cidade Universitária, 50732-970 Recife, PE, Brasil
acs@cin.ufpe.br

Abstract: Spatial data base systems (SDBSs) deal with data that are special in nature and size. Thus, the technologies developed for conventional data base systems such as physical design, access methods, query optimizers and languages, have to be modified in order to satisfy the needs of a SDBS. These modifications, embedded in several SDBSs, or being proposed by research projects, need to be evaluated. Our research focuses specifically multidimensional access methods (MAMs), which are very important structures used to diminish the execution time of search operations in a SDBS. In this paper, we discuss two important factors that affect the performance of MAMs: degree of overlap between non-zero size data objects and data selectivity.

1 Introdução

A manipulação de dados espaciais, tais como pontos, linhas, polígonos e sólidos tridimensionais, é um requisito fundamental para inúmeras aplicações atuais, as quais incluem aplicações geográficas de gerenciamento de serviços de utilidade pública (telefonia, eletricidade, água e esgoto), aplicações de planejamento e administração cadastral urbano e rural, aplicações ambientais, aplicações CAD/CAM (projeto e fabricação auxiliados por computador), aplicações cartográficas e aplicações de biotecnologia.

O gerenciamento em memória secundária de dados que possuem extensão e/ou localização no espaço é tarefa específica de uma classe particular de sistemas de banco de dados (SBDs), denominada de SBDs espaciais (SBDEs). Sistemas de banco de dados espaciais estendem as funcionalidade típicas de gerenciamento de dados alfanuméricos em disco, oferecendo suporte integrado a tipos de dados espaciais em seu modelo de dados, linguagem de consulta e implementação física.

Dentre as várias estruturas utilizadas por um SBDE, métodos de acesso multidimensionais (MAMs) destacam-se por permitirem que o desempenho na recuperação de dados seja melhorado significativamente. Estes métodos, também conhecidos como mecanismos de indexação espacial, organizam o espaço multidimensional e os objetos contidos neste espaço de tal forma que somente algumas partes do espaço e um subconjunto dos objetos espaciais armazenados sejam considerados para responder uma dada consulta espacial. Esta organização é fundamental, devido ao rápido crescimento do volume de dados armazenado em SBDEs. Os bancos de dados do sistema *Earth Observing System* da NASA e do projeto SEQUOIA 2000, por exemplo, armazenarão respectivamente, quando finalizada a fase de coleta de dados, 10 Pbytes (10^{16}) e 100 Tbytes (10^{14}) de dados.

Após duas décadas de pesquisa, diversos MAMs foram desenvolvidos, os quais são voltados principalmente para a indexação de pontos e retângulos. Gaede e Günther [8] realizaram uma extensiva revisão bibliográfica dos métodos de acesso propostos até 1996. Este levantamento revelou a existência de aproximadamente 50 métodos de acesso. Infelizmente, esforço compatível não tem sido verificado na comparação de desempenho de MAMs. Poucos trabalhos efetuaram testes de desempenho comparativos, sendo que estes trabalhos abrangeram apenas um pequeno subconjunto dos métodos de acesso, com ênfase na família R-tree. Ademais, alguns trabalhos mostram-se limitados na investigação de certos fatores determinantes de desempenho, isto é, fatores que influenciam diretamente o desempenho dos MAMs no suporte a consultas espaciais e no suporte a operações de inserção, remoção e modificação de dados.

A pesquisa sendo desenvolvida pelos autores deste artigo está direcionada para a análise de desempenho de métodos de acesso multidimensionais através do uso da técnica experimental de *benchmark* de banco de dados. Neste sentido, primeiramente foram identificados os principais fatores determinantes de desempenho (Tabela 1). Em seguida, foi feito um levantamento de trabalhos anteriores voltados à investigação e comparação da variação do desempenho de MAMs. Após este levantamento, foi realizada uma análise crítica e comparativa dos diversos trabalhos correlatos em função de cada um dos fatores determinantes de desempenho. Vale destacar que a referida análise crítica visou identificar problemas e limitações nas metodologias usadas para investigar a influência dos

fatores na eficiência dos MAMs. Tanto o levantamento de trabalhos correlatos, quanto a análise crítica e comparativa podem ser encontrados em Ciferri e Salgado [5].

A análise de desempenho pretende abranger diversos MAMs, tais como R-tree com algoritmos de particionamento de nó quadrático e linear, R-tree Greene, R^+-tree, R^*-tree, Hilbert R-tree, cell tree with oversize shelves, PMR-quadtree e M-tree. Um ambiente de teste deve incluir: (1) um conjunto representativo de arquivos de dados reais e sintéticos (isto é, gerados artificialmente); (2) uma interface funcional que permita a comunicação entre usuários finais e a ferramenta de *benchmark* e consequentemente torne possível a parametrização dos fatores determinantes de desempenho e (3) mecanismos para a visualização estatística dos resultados de desempenho coletados. Nosso ambiente de teste, além de atender a estes requisitos, medirá o desempenho segundo diferentes perspectivas, ou seja, através da utilização de diferentes tipos de medidas de desempenho (acessos a disco, operações geométricas, tempo de utilização da UCP, espaço de armazenamento, tempo total gasto, etc.).

No estágio atual de nossa pesquisa, somente o fator distribuição espacial dos dados foi investigado para um subconjunto dos métodos de acesso anteriormente citados (métodos de acesso da família R-tree). Para isto, foi proposta uma metodologia que permite a geração de um conjunto de distribuições de dados com diferentes características, as quais tornam possível que a influência do fator distribuição espacial dos dados seja analisada sob perspectivas distintas, desde uma reduzida até uma acentuada influência no desempenho dos MAMs. Como resultado, confirmou-se que a distribuição espacial dos dados exerce grande influência no desempenho tanto absoluto quanto relativo de MAMs. Ademais, em especial, foram verificadas novas relações de desempenho, algumas das quais contrariam resultados de desempenho e conclusões obtidas em trabalhos anteriores, tais como em Cox Júnior [7] e em Carneiro [3]. Maiores detalhes sobre a investigação do fator distribuição espacial dos dados podem ser encontrados em [6].

Este artigo discute especificamente aspectos relacionados com os fatores determinantes de desempenho "grau de sobreposição entre objetos espaciais de dimensão não-zero" e "seletividade dos dados". Estes dois fatores, apesar de importantes, tem recebido um tratamento superficial em investigações anteriores. A próxima seção discute aspectos relacionados com o grau de sobreposição entre objetos espaciais de dimensão não-zero. O controle da seletividade dos dados e as diversas aplicações deste controle são discutidos na seção 3. Já a seção 4 apresenta brevemente algumas características de trabalhos correlatos. Por fim, a seção 5 apresenta as conclusões.

classes de fatores	fatores determinantes de desempenho
relacionados aos dados	tipo de dado, distribuição espacial dos dados, volume (quantidade) e escalabilidade, tamanho e formato de objetos espaciais de dimensão não-zero, grau de sobreposição entre objetos espaciais de dimensão não-zero e seqüência de inserção dos dados.
relacionados à carga de trabalho	tipo de consulta (ex.: *nearest neighbour query*), características associadas a parâmetros de consulta (tais como tamanho, formato e distribuição espacial dos retângulos relativos às janelas de consulta de *range queries* ou distribuição espacial dos pontos base de consultas do tipo *point query*), dinâmica dos dados (operações de inserção, remoção e modificação) e seletividade dos dados.
relacionados ao gerenciamento de *buffer-pool*	política de gerenciamento, tamanho total do *buffer* e tamanho da página de disco. A variação deste último fator conduz, para um mesmo MAM, a diferentes agrupamentos de objetos. Já os demais fatores determinam quais e quantos objetos estarão presentes em memória principal, requisitos fundamentais para otimizar os acessos a disco.
relacionados a parâmetros da estrutura	a variação de alguns parâmetros, tal como o número mínimo de entradas por nó para os MAMs R-tree e R^*-tree, permite um ajuste diferenciado na estrutura de dados e por conseguinte um melhor ou pior desempenho.

Tabela 1 Fatores determinantes de desempenho de métodos de acesso multidimensionais

2 Grau de Sobreposição entre Objetos Espaciais de Dimensão não-zero

Diversos trabalhos têm destacado a influência da sobreposição entre objetos espaciais no desempenho de MAMs [8, 9, 14]. Greene [9], por exemplo, concluiu que o método R^+-tree produz um melhor desempenho quando ocorrem menos sobreposições entre objetos espaciais de dimensão não-zero. Já Kriegel *et all* [14], em suas considerações finais, destacam algumas limitações presentes na avaliação de desempenho realizada pelos próprios autores, dentre as quais encontra-se a ausência de testes de desempenho que investigassem os efeitos da variação do grau de sobreposição. A caracterização exata do termo "grau de sobreposi-

ção", no entanto, não foi efetuada por Kriegel *et all*, nem pelos demais trabalhos correlatos estudados.

O grau de sobreposição, segundo o entendimento dos autores deste trabalho, pode ser definido a partir de duas medidas. A primeira consiste na média do número de objetos intersectados por cada objeto espacial armazenado em um arquivo de dados. Outra medida possível para contabilizar o grau de sobreposição é baseada na extensão espacial envolvida. Para retângulos no espaço bidimensional, por exemplo, esta medida deve ser calculada em dois passos. Inicialmente, para cada objeto espacial deve-se calcular a fração de área envolvida com sobreposições, sendo que esta fração deve ser quantificada através de um percentual de área relativo à área total do *extent*. Por fim, calcula-se a média das frações. Um alto grau de sobreposição corresponde, portanto, a um grande número de objetos intersectados e a uma vasta área de sobreposição. A Figura 1 ilustra um exemplo no qual um dado objeto espacial intersecta três outros objetos, sendo que a fração de área envolvida com sobreposições está em cinza.

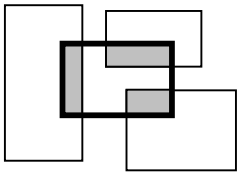

Figura 1 Exemplo de sobreposição

Ao invés de controlar o grau de sobreposição segundo as medidas acima citadas, por exemplo na geração de dados sintéticos, alguns trabalhos correlatos apresentaram apenas uma estimativa do grau de sobreposição médio presente nos vários arquivos de dados através da medida "densidade de sobreposição" [1, 9, 11, 13]. Esta medida indica quantas vezes o *extent* é coberto pela *coverage* (área total) de todos os objetos espaciais, e pode ser calculada pela Equação 1, onde *n* corresponde ao número de objetos espaciais. A *coverage* de todos os objetos espaciais, de modo equivalente, também pode ser calculada através da expressão *n* * tamanho médio dos objetos.

$$Densidade = \frac{\sum_{i=1}^{n} tamanho\_objeto(i)}{tamanho\_extent}$$

Equação 1 Densidade de sobreposição

A medida densidade de sobreposição, segundo alguns pesquisadores [10, 15], é um bom indicador do grau de sobreposição, sendo que valores crescentes de densidade tendem a indicar um maior grau de sobreposição. Neste sentido, Günther *et all* [11] apresenta alguns exemplos com densidades distintas, sendo que para uma densidade de sobreposição de 100% foi verificado que sobreposições entre objetos espaciais ocorriam, mas em um grau não muito elevado, enquanto que uma densidade de 5% tornava praticamente inexistente a ocorrência de sobreposições entre objetos espaciais. A Tabela 2 apresenta as densidades consideradas em alguns trabalhos.

trabalhos correlatos voltados à análise de desempenho de MAMs	**densidade de sobreposição**
Greene [9]	0,0001; 0,01; 1 e 10
Beckmann *et all* [1]	2; 2,5; 8; 10 e 11,2
Kamel e Faloutsos [13]	entre 0,029 e 1
Günther *et all* [11]	0,05; 0,3 e 1

Tabela 2 trabalhos correlatos

A medida densidade de sobreposição, entretanto, possui como grande limitação o fato de não ser muito precisa, sendo que situações completamente antagônicas podem apresentar o mesmo valor de densidade. Como exemplo, a Figura 2 ilustra o conteúdo de dois arquivos de dados com a mesma densidade de sobreposição, mas claramente contendo diferentes graus de sobreposição. Vale destacar que, em um destes arquivos, propositadamente, não ocorre nenhuma sobreposição. Desta forma, pode-se concluir que o uso isolado da medida densidade de sobreposição na investigação da variação do desempenho de MAMs em função do grau de sobreposição não é apropriado, sendo necessário o uso de um conjunto alternativo de medidas que caracterize de uma forma mais precisa o grau de sobreposição.

Ainda há espaço, portanto, para novas investigações sobre a influência da sobreposição entre objetos espaciais no desempenho de MAMs, a qual deve ser conduzida de um modo controlado e exaustivo. A obtenção de diferentes valores para o grau de sobreposição pode ser alcançada através da variação de outros fatores determinantes de desempenho, os quais influenciam indiretamente no grau de sobreposição. Cox Júnior [7] observou que conjuntos de dados com objetos grandes, por exemplo, tendem a gerar sobreposições e provocam alterações de desempenho distintas por parte dos métodos de acesso. A distribuição espacial dos dados, por sua vez, também influencia no grau de sobreposição, sendo um bom exemplo ilustrado na Figura 2. Ooi [15] verificou que quando os objetos estão distribuídos de uma forma mais uniforme, o grau de sobreposição é menor.

Vale destacar, entretanto, que a realização de testes que investigam isoladamente a variação do desempenho em função da distribuição espacial dos dados ou em função do tamanho dos objetos espaciais não proporcionam qualquer conclusão independente com relação à sobreposição entre objetos espaciais, apesar destes testes também conduzirem à variação do grau de sobreposição.

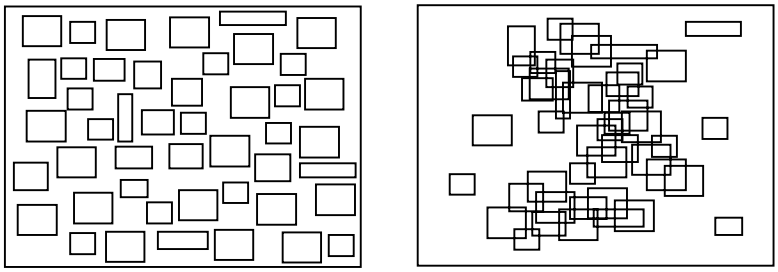

Figura 2 Exemplo de arquivos com a mesma densidade de sobreposição, mas contendo diferentes graus de sobreposição

A variação da distribuição espacial dos dados para dois arquivos de um mesmo tipo[1], por exemplo, em geral conduz à variação do grau de sobreposição, tal como um arquivo de dados com distribuição altamente correlacionada apresentando um grau de sobreposição cinqüenta vezes maior do que um arquivo com distribuição uniforme. Neste sentido, pode-se concluir, incorretamente, que um método de acesso multidimensional *A* apresenta um desempenho superior ao método de acesso *B* quando o grau de sobreposição é alto, ao passo que para um baixo grau de sobreposição o método de acesso *A* apresenta os piores resultados. Neste caso específico, entretanto, o desempenho é influenciado prioritariamente pela distribuição espacial dos dados, sendo que a influência do grau de sobreposição pode ser considerada como secundária. Observação similar pode ser feita no caso da variação do tamanho dos objetos, com destaque para a manutenção da distribuição espacial dos dados.

Para verificar de fato qual a influência da sobreposição entre objetos espaciais no desempenho de métodos de acesso multidimensionais deve-se variar conjuntamente ambos os fatores distribuição espacial dos dados e tamanho dos objetos espaciais, os quais devem ser usados para controlar o grau de sobreposição. Uma possível estratégia consiste da geração de um conjunto de pares de arquivos, sendo que cada par possui um grau de sobreposição distinto, mas os arquivos de um dado par possuem o mesmo grau de sobreposição (ou muito próximos) e são formados por diferentes distribuições de dados e tamanhos de objetos. Como exemplo, pode-se citar a geração do par de arquivos *A* e *B*, com o arquivo *A* constituído de objetos pequenos sob distribuição altamente correlacionada e o arquivo *B* formado por objetos grandes distribuídos de modo uniforme no espaço, mas ambos arquivos apresentando o mesmo grau de sobreposição.

Caso o desempenho, para um determinado método de acesso, seja semelhante para os arquivos de dados de um mesmo par e significativamente diferente entre os vários pares de arquivos[2], comprova-se que o desempenho foi influenciado basicamente pela sobreposição dos dados (e independente, de certo modo, da distribuição espacial dos dados e do tamanho dos objetos espaciais). Por outro lado, a constatação freqüente, para diferentes métodos de acesso, de um desempenho muito diferenciado para arquivos de dados de um mesmo par ou de um desempenho similar para pares com graus de sobreposição muito distintos (ausência, baixo, médio e alto, por exemplo), pode indicar que o grau de sobreposição não constitui em um fator determinante de desempenho para MAMs.

3 Controle da Seletividade dos Dados

A seletividade dos dados corresponde ao percentual de dados recuperados por uma consulta com relação ao conjunto completo de possíveis respostas à consulta. No caso de consultas espaciais de seleção, este percentual é calculado, portanto, com base na quantidade total de objetos espaciais armazenados no arquivo de dados sobre o qual a consulta foi executada, enquanto para consultas do tipo junção espacial, este percentual é calculado com base no número de tuplas do produto cartesiano de dois arquivos de dados.

O controle da seletividade dos dados, isto é, a possibilidade de se recuperar somente uma certa percentagem dos dados, como exemplo 1%, 5%, 15% ou 30%, é muito importante, dado que uma consulta pode gerar um desempenho não proporcional ao percentual de dados recuperados [4]. Uma consulta, logicamente, pode gerar um desempenho distinto em função do percentual de dados recuperados, ocorrendo de modo geral uma diminuição no desempenho à medida que a seletividade dos dados aumenta, ou seja, à medida que cresce o volume de dados recuperados. Entretanto, apesar da diferença de desempenho obtida com a utilização de graus de seletividade distintos, espera-se que a razão entre os desempenhos seja proporcional à razão entre os graus de seletividade.

[1] arquivos com características equivalentes quanto a tipo de dado, volume, tamanho e formato dos objetos espaciais.

[2] deve-se usar graus de sobreposição com certa diferença, de modo que a variação não reflita uma variação contínua.

Como exemplo, se a recuperação de 10% dos objetos espaciais armazenados em um arquivo de dados gastar 5 segundos, a recuperação de 30% e 50% dos objetos espaciais deve gastar respectivamente em torno de 15 e 25 segundos. Quando a recuperação de 30% dos objetos espaciais consumir, por exemplo, 30 segundos, verifica-se um desempenho não proporcional ao percentual de dados recuperados. Em especial, Günther *et all* [11] destaca que a seletividade dos dados constitui em um excelente indicador de variações no desempenho relativo. Para sistemas gerenciadores de banco de dados relacionais, o estudo da variação do desempenho em função do percentual de dados recuperados por uma consulta tem sido largamente realizado, com destaque para o uso do *benchmark* de Wisconsin.

A investigação da variação do desempenho de métodos de acesso em função do percentual de dados recuperados por uma consulta é tanto viável quanto particularmente interessante. Viável no sentido que alguns trabalhos correlatos já constataram, de forma indireta, que o aumento do grau de seletividade dos dados pode alterar o desempenho relativo dos métodos de acesso, assim como pode alterar o desempenho absoluto de um MAM. Esta constatação foi feita a partir da variação do tamanho da janela de consulta de *intersection range queries*. O crescimento do tamanho da janela de consulta implica, indiretamente, em um aumento do grau de seletividade para a maioria absoluta das situações[3]. Como exemplo, Carneiro [3] verificou que a diferença de desempenho entre as variantes da R^*-tree e da R-tree tende a cair com o aumento do tamanho da janela de consulta. A simples variação do tamanho da janela de consulta não permite, entretanto, um controle preciso do grau de seletividade. Para isto, adicionalmente é necessário uma distribuição sistemática dos dados e das janelas de consulta.

Dentre os trabalhos correlatos revisados, apenas dois trabalhos consideraram de algum modo o fator seletividade dos dados nos testes de desempenho. O trabalho de Guttman [12] não investigou a variação do desempenho do método de acesso R-tree em função do percentual de dados recuperados por uma consulta, e por conseguinte não controlou a variação do grau de seletividade dos dados, limitando-se apenas a constatar que a execução das consultas, no caso *intersection range query*, conduzia a uma seletividade entre 3% e 6% do total de objetos espaciais armazenados nos arquivos de dados.

Já o trabalho de Günther *et all* [11] destaca-se por ter investigado a variação do desempenho de estratégias distintas de processamento de junção espacial em função do percentual de dados recuperados por vários subtipos de junção espacial, cada qual com um predicado específico, tais como *distance minor-equal*, *intersection* e *direction northwest*. Apesar do enfoque dado aos testes de desempenho não ter sido direcionado para métodos de acesso, a existência de estratégias fundamentadas no uso de MAMs, tais como *scan-and-index* e *synchronized tree traversal*, permitiu uma avaliação do método de acesso quadtree, escolhido para dar suporte a estas estratégias.

Em especial, verifica-se que ainda existe espaço para se realizar novas pesquisas com métodos de acesso multidimensionais envolvendo o fator seletividade dos dados. Neste sentido, pode-se realizar testes que:

- investiguem de forma precisa e controlada o comportamento de métodos de acesso específicos, tais como os MAMs da família R-tree, frente a diferentes graus de seletividade dos dados. Em particular, pode-se verificar se o desempenho do método de acesso em questão é proporcional ou não ao percentual de dados recuperados por uma consulta. Quanto ao tipo de consulta, pode-se destacar consultas do tipo *range query*, uma vez que este tipo de consulta tem sido amplamente investigado na literatura e possui grande popularidade, por exemplo, entre usuários de aplicações georeferenciadas;
- investiguem a variação do desempenho relativo de MAMs em função do percentual de dados recuperados por uma consulta. Para métodos de acesso com estrutura na forma de uma árvore de múltiplos caminhos (*multiway tree*), por exemplo, o agrupamento de objetos espaciais em nós folhas deve ser feito de tal maneira que um baixo grau de seletividade (tal como 1% do número total de objetos espaciais indexados) conduza ao acesso a um número reduzido de nós, especialmente de nós folhas. Caso contrário, pode ocorrer uma degradação excessiva no desempenho. Já para um alto grau de seletividade, para o qual ocorre a recuperação de grande parte dos objetos indexados (como exemplo, 75%), espera-se o acesso a uma grande quantidade de nós, sendo a eficiência do agrupamento de objetos espaciais menos importante neste caso. Assim, pode-se verificar o desempenho relativo de vários MAMs tanto para um baixo grau de seletividade dos dados, para o qual é esperado uma diferença significativa no desempenho, quanto para um alto grau de seletividade dos dados, para o qual espera-se desempenhos semelhantes;
- investiguem a influência que a distribuição espacial dos dados exerce na seletividade dos dados. Isto pode ser conseguido através da recuperação de um mesmo percentual de dados para regiões distintas do *extent*. Como resultado, poderá ocorrer duas possíveis situações. Na primeira situação, o desempenho é determinado predominantemente pelo grau de seletividade e portanto praticamente o mesmo

[3] considerando que as janelas possuem o mesmo formato e distribuição espacial do centro, diferindo somente com relação ao tamanho. Assim, a seletividade dos dados para duas janelas com mesmo centro, na pior hipótese, será igual.

quando consideradas regiões distintas do *extent*. Já na segunda situação, verifica-se que o desempenho difere de forma significativa de uma região para outra, apesar do percentual de dados recuperados pela consulta ser o mesmo. Neste caso, em particular, pode-se dizer que o desempenho de um método de acesso, para um dado grau de seletividade, é influenciado pela distribuição espacial dos dados;

A seletividade dos dados também destaca-se pela função que exerce no suporte à escolha de um plano de execução por parte de um otimizador de consultas. Este fator é comumente estimado para determinar a escolha de planos de execução que usam índices convencionais ou métodos de acesso multidimensionais. Dependendo da seletividade estimada dos dados, o otimizador de consultas de um sistema gerenciador de banco de dados (SGBD) pode escolher, por exemplo, entre o uso de um índice ou a realização de uma busca seqüencial. Em geral, para graus de seletividade muito altos, o uso de índices não é recomendado, uma vez que a sobrecarga de execução causada por estas estruturas supera qualquer ganho de desempenho potencial. Neste caso, a realização de uma busca seqüencial gera um melhor desempenho.

A escolha de uma estratégia para processamento de consultas em um SGBD espacial deve considerar ambos os predicados espacial e convencional. Neste sentido, o otimizador de consultas deve primeiramente gerar diferentes planos de execução, os quais definem a ordem de processamento das diversas subconsultas associadas aos predicados, sendo que em alguns casos a ordem de processamento restringe o conjunto de dados no qual uma subconsulta posterior irá atuar. Com base em uma estimativa de custo associado a cada plano, o otimizador de consultas deve então escolher o plano com menor custo. Câmara, *et all* [2] apresenta e discute diferentes planos de execução para consultas com ambos os tipos de predicado espacial e convencional, com o predicado espacial abrangendo especificamente consultas do tipo *region query* e junção espacial.

Há casos, no entanto, que uma consulta é constituída apenas por um predicado espacial. Um típico exemplo é "selecione todos os municípios que estão localizados na zona da mata de Pernambuco" (*containment region query*). Nestes casos, a escolha do otimizador de consulta resume-se basicamente a dois planos: (1) o uso de um método de acesso multidimensional ou (2) a varredura completa do arquivo de dados.

No **primeiro plano**, assume-se, claro, que o respectivo arquivo de dados encontra-se indexado. Vale destacar que este plano não se restringe unicamente ao uso de um método de acesso, sendo adicionalmente necessário uma fase posterior de processamento. Tal necessidade é conseqüência direta do fato que métodos de acesso representam a geometria de objetos espaciais de dimensão não-zero, como exemplo uma linha ou um polígono, através de uma aproximação, visando diminuir tanto o custo de armazenamento quanto o custo de processamento para se determinar a satisfação dos relacionamentos espaciais. As aproximações, tal como MBB, garantem que nenhum dos objetos espaciais que satisfazem o relacionamento em questão seja desconsiderado na resposta da consulta, mas em contrapartida permitem a recuperação de objetos espaciais que não satisfazem o relacionamento espacial (chamados de falsos candidatos), uma vez que a determinação do relacionamento pode ser decidida com base em um espaço, contido dentro da aproximação, que não faz parte da geometria do objeto (*dead space* – Figura 3). Assim, pode-se dizer que MAMs são imprecisos, no sentido que estes não retornam a resposta final e exata das consultas, retornando ao invés disto um superconjunto de candidatos. Na literatura, esta fase é comumente denominada de filtragem, uma vez que um MAM apenas filtra (ou descarta) os objetos que certamente não satisfazem o relacionamento espacial em questão.

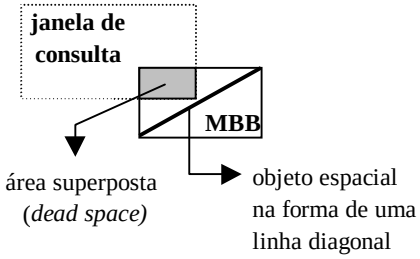

Figura 3 Recuperação de um falso candidato

Devido à presença de falsos candidatos no conjunto de objetos selecionados, é necessário uma fase posterior de processamento, chamada de refinamento, na qual é verificado, para cada objeto espacial candidato, se o relacionamento espacial é satisfeito com relação à geometria exata do objeto, sendo descartados quaisquer falsos candidatos. A fase de refinamento é altamente custosa, pois requer tanto o acesso à geometria exata dos objetos espaciais, e portanto a transferência para memória principal de uma quantidade relativamente grande de dados, quanto a realização de cálculos geométricos complexos para se determinar a satisfação do relacionamento espacial. O custo do primeiro plano deve ser calculado, portanto, a partir dos custos de ambas as fases de filtragem e refinamento.

Já o **segundo plano** resume-se à aplicação direta da fase de refinamento para todos os objetos espaciais armazenados no arquivo de dados (estratégia equivalente a uma busca seqüencial). Desta forma, para cada objeto espacial é necessário o acesso à sua geometria exata e a realização

subsequente de cálculos geométricos sobre esta geometria para se determinar a satisfação do relacionamento espacial.

A escolha por parte do otimizador de consultas entre os dois planos de execução anteriormente citados é altamente influenciada pela seletividade estimada dos dados, a qual exerce grande peso na determinação do custo dos planos. A escolha do primeiro plano, e portanto o uso de um MAM, é baseado nas seguintes premissas: (1) a fase de refinamento é altamente custosa e deve ser aplicada a uma quantidade extremamente reduzida de objetos espaciais; (2) a fase de filtragem, por sua vez, é várias vezes menos custosa do que a fase de refinamento e assim, deve ser usada para restringir o número de objetos (candidatos) a serem analisados na fase de refinamento. Desta forma, a fase de filtragem serve para diminuir o custo da fase de refinamento, a qual basicamente determina o desempenho de uma consulta.

A escolha do primeiro plano é válida, no entanto, somente se a fase de filtragem realmente reduzir drasticamente o número de objetos a serem analisados na fase de refinamento, ou seja, se o grau de seletividade for baixo. Neste caso, a soma dos custos de ambas as fases de filtragem e refinamento tende a ser menor do que o custo resultante da aplicação direta da fase de refinamento para todos os objetos espaciais armazenados no arquivo de dados (segundo plano), devido principalmente à diferença significativa nos custos da fase de refinamento de ambos os planos. Por outro lado, se o grau de seletividade for muito alto, esta diferença torna-se pequena, podendo até ser superada pelo custo associado à fase de filtragem do primeiro plano de execução, o que resultaria em um melhor desempenho para o segundo plano.

Visando proporcionar heurísticas confiáveis para otimizadores de consulta espaciais, surge então a necessidade de se determinar para quais graus de seletividade dos dados a utilização de um método de acesso multidimensional não torna-se desvantajosa. Para isto, deve-se investigar o desempenho de ambos os planos de execução segundo diferentes graus de seletividade dos dados, desde graus reduzidos (como exemplo 1%) até graus elevados de seletividade (tais como, 80% ou 90%). Vale destacar, entretanto, que o ponto de equilíbrio (ou seja, o grau de seletividade dos dados segundo o qual ambos os planos proporcionam o mesmo desempenho no processamento de um dado tipo de consulta espacial) varia em função do método de acesso escolhido para dar suporte ao primeiro plano, sendo a investigação voltada a um tipo específico de MAM. Neste sentido, a análise de diversos métodos de acesso é particularmente interessante.

De modo análogo à discussão realizada para consultas espaciais de seleção, pode-se também investigar sobre o ponto de equilíbrio no processamento de consultas do tipo junção espacial, ou seja, identificar o grau de seletividade segundo o qual o uso de uma estratégia baseada em MAM, tal como *scan-and-index*, produz o mesmo desempenho que o processamento completo do produto cartesiano de dois conjuntos de dados espaciais.

4 Trabalhos Correlatos

Guttman [12] realizou testes de desempenho exclusivamente voltados para a validação e ajuste da estrutura R-tree, sendo considerados diferentes valores para os parâmetros *M* e *m* (número máximo e mínimo de entradas por nó, respectivamente) e diferentes algoritmos de particionamento de nó (exaustivo, quadrático e linear).

Kriegel *et all* [14] realizaram testes específicos tanto para métodos de acesso a pontos (2-level grid file, BANG file, hB-tree e BUDDY hash tree), quanto para métodos de acesso espaciais (R-tree e PLOP hashing). Esse trabalho possui como característica principal o uso de diversos tipos de distribuição de dados, com destaque para distribuições altamente correlacionadas.

O principal diferencial dos testes realizados por Greene [9] está relacionado com o fato de que os MAMs analisados (R-tree, R^+-tree, K-D-B-tree e 2D-ISAM) foram implementados no topo do SGDB POSTGRES através do uso de procedimentos especiais com acesso as funções da camada de abstração de métodos de acesso deste SGBD.

Beckmann *et all* [1] realizaram testes para ajustar, validar e comparar a estrutura proposta R^*-tree com outros métodos de acesso, a saber: R-tree com algoritmos de particionamento de nó quadrático e linear, R-tree Greene e 2-level grid file. A execução dos testes foi efetuada em três etapas distintas. A primeira foi voltada à análise de desempenho dos MAMs da família R-tree no suporte a retângulos, mas considerando apenas consultas espaciais de seleção. Já a segunda gerou configurações de teste para consultas do tipo junção espacial. Por fim, a última etapa foi voltada à análise de desempenho no suporte a pontos.

Günther e Bilmes [10] realizaram testes de desempenho para comparar um tipo alternativo de MAM, chamado cell tree, com os métodos de acesso R-tree Greene e R^+-tree. A principal característica da cell tree é que esta estrutura provê suporte direto a objetos espaciais de dimensão não-zero com geometria arbitrária, em oposição ao suporte restrito e único de retângulos.

A estratégia utilizada nos testes realizados por Cox Júnior [7] é baseada na geração sintética de arquivos de dados e de consultas a partir de um conjunto de fatores determinantes de desempenho (tipo de dado, tipo de consulta, tamanho dos objetos, distribuição espacial e dinâmica dos dados). Em particular, foram analisados os métodos R-tree, R-tree Greene, R^+-tree, R^*-tree e uma variante desta

sem a rotina de reinserção de entradas. Carneiro [3] adaptou e estendeu a estratégia proposta em [7] visando o uso de um conjunto de dados reais representativos para aplicações de gerenciamento de serviços de utilidade pública.

Ooi [15] realizou testes de desempenho voltados para a validação, o ajuste e a comparação de desempenho da estrutura spatial kd-tree (skd-tree), para ambas versões estática e dinâmica. Na comparação, foram investigados os métodos de acesso multidimensionais R-tree, packed R-tree, 4d-tree e mkd-tree.

Kamel e Faloutsos [13] realizaram experimentos para comparar o método de acesso proposto Hilbert R-tree com relação aos métodos de acesso R-tree e R^*-tree. Na comparação dos métodos foi utilizada apenas uma única medida, número de acessos a disco, a qual foi usada principalmente na coleta de resultados de desempenho relativos à consultas do tipo *intersection range query*.

Por fim, Günther *et all* [11] investigaram o desempenho das estratégias *nested loop*, *synchronized tree traversal* e *scan-and-index* no suporte a consultas do tipo *intersection spatial join*, *containment spatial join*, *distance minor-equal spatial join*, *direction northwest spatial join* e *enclosure spatial join*.

5 Conclusão

Este artigo descreveu o atual estágio da pesquisa sendo desenvolvida pelos autores com relação a métodos de acesso multidimensionais. Ademais, também foram discutidos aspectos relacionados com os fatores determinantes de desempenho "grau de sobreposição entre objetos espaciais de dimensão não-zero" e "seletividade dos dados".

Para os fatores acima citados, verificou-se a necessidade de novos testes de desempenho, os quais devem incorporar novas metodologias para investigar a influência destes fatores na eficiência dos MAMs. Para o fator "grau de sobreposição entre objetos espaciais de dimensão não-zero", uma possível estratégia foi apresentada na seção 2. Já o controle da seletividade dos dados e as diversas aplicações deste controle foram discutidos na seção 3. Vale destacar que a pesquisa descrita neste artigo representa um primeiro passo para a definição de um *benchmark* de banco de dados voltado à análise de desempenho de MAMs.

Referências Bibliográficas

[1] Beckmann, N., Kriegel, H.-P., Schneider, R., Seeger, B The R*-Tree: An Efficient and Robust Access Method for Points and Rectangles. In *Proc. 1990 ACM SIGMOD Conference*, pp. 322-331, 1990.

[2] Câmara, G., Casanova, M.A., Hemerly, A.S., Magalhães, G.C., Medeiros, C.M.B. *Anatomia de Sistemas de Informação Geográfica*. 1996. 193pp.

[3] Carneiro, A.P. Análise de Desempenho de Métodos de Acesso Espaciais Baseada em um Banco de Dados Real. Master's thesis, IC, Unicamp, 1998.

[4] Ciferri, R.R. Um Benchmark voltado à Análise de Desempenho de Sistemas de Informações Geográficas. Master's thesis, Unicamp, 1995.

[5] Ciferri, R.R, Salgado, A.C. Análise de Desempenho de Métodos de Acesso Multidimensionais: Estado da Arte e Direções Futuras. Technical report, CIN/UFPE, 2000. 105 pp.

[6] Ciferri, R.R, Salgado, A.C., Cortez, S.S. Investigando a Variação do Desempenho de Métodos de Acesso Multidimensionais em Função da Distribuição Espacial dos Dados. Submetido para julgamento no SBBD'2000.

[7] Cox Junior, F.S. Análise de Métodos de Acesso a Dados Espaciais Aplicados a Sistemas Gerenciadores de Banco de Dados. Master's thesis, DCC, Unicamp, 1991.

[8] Gaede, V., Günther, O. Multidimensional Access Methods. *ACM Computing Surveys*, 30(2):170-231,1998.

[9] Greene, D. An Implementation and Performance Analysis of Spatial Data Access Methods. In *Proc. 5th IEEE ICDE*, pp. 606-615, USA, 1989.

[10] Günther, O., Bilmes, J. Tree-based Access Methods for Spatial Databases: Implementation and Performance Evaluation. *IEEE TKDE*, 3(3):342-356, September 1991.

[11] Günther, O., Oria, V., Picouet, P., Saglio, J.-M., Scholl, M. Benchmarking Spatial Joins À La Carte. In *Proc. 10th Conference on Scientific and Statistical Database*, pp. 32-41, Italy, 1998.

[12] Guttman, A. R-Trees: A Dynamic Index Structure for Spatial Searching. In *Proc. 1984 ACM SIGMOD Conference*, pp. 47-57, USA, June 1984.

[13] Kamel, I., Faloutsos, C. Hilbert R-tree: An Improved R-tree using Fractals. In *Proc. 20th VLDB Conference*, pp. 500-509, Chile, 1994.

[14] Kriegel, H.-P., Schiwietz, M., Schneider, R., Seeger, B. Performance Comparison of Point and Spatial Access Methods. In *Design and Implementation of Large Spatial Databases*, v. 409 of *LNCS*, pp. 89-114. 1989.

[15] Ooi, B.C. Efficient Query Processing in Geographic Information Systems. In Goos, G., Hartmanis, J., editors, volume 471 of *LNCS*, chapter 4, pp. 116-135. 1990.