Uso de Lógica Nebulosa na Construção e na Utilização da Árvore Métrica Slim-tree

Cláudio Haruo Yamamoto[1]

Mauro Biajiz (orientador)[2]

Departamento de Computação
Universidade Federal de São Carlos
Rodovia Washington Luis, Km 235, Caixa Postal 676
CEP: 13565-905, São Carlos, SP, Brasil
Tel: (16) 260-8232, Fax: (16) 260-8233
{charuo[1], mauro[2]}@dc.ufscar.br

Resumo

O objetivo deste trabalho é construir uma árvore métrica utilizando conceitos da Lógica Nebulosa. A nova árvore terá como base a Slim-tree e receberá o nome de Fuzzy Slim-tree. O desempenho de árvores métricas está relacionado ao número de páginas acessadas, à quantidade de cálculos de distância efetuados e à utilização de espaço de armazenamento. A aplicação de conceitos da Lógica Nebulosa visa melhorar o desempenho dos algoritmos associados à Slim-tree, melhorando a distribuição dos objetos dentro dos nós e diminuindo o número de sobreposições dos mesmos. Para esse fim, está sendo proposta uma nova estrutura que incorpore conceitos da Lógica Nebulosa e que melhore assim os algoritmos de inserção, divisão de nó e escolha dos representativos, otimização e consulta associados à Slim-tree.

Palavras-chave: indexação métrica, métodos de acesso métricos, árvores métricas, estruturas de dados, lógica nebulosa.

1. Introdução

Desde os anos 70, várias estruturas de indexação têm sido criadas para uma gama diferente de aplicações que utilizam dados representados em espaços multidimensionais (que são representados em domínios espaciais n-dimensionais) [9] e métricos (que são representados em espaços determinados por métricas pré-definidas e que obedecem às restrições destes espaços) [12]. São aplicações para estas estruturas de indexação o *Computer Aided Design* (*CAD*), os Sistemas de Informação Geográficos (*GIS*), o processamento de imagens médicas, dentre muitas outras. As estruturas de indexação podem ser divididas em quatro grandes categorias: as estruturas básicas, os Métodos de Acesso Espacial Puntuais (MAEP), os Métodos de Acesso Espacial Não-Puntuais (MAENP) e os Métodos de Acesso Métricos (MAM). O primeiro grupo compreende as estruturas a partir das quais as outras estruturas foram derivadas. As técnicas de *hashing* e a B-tree, dentre outras são exemplos deste grupo. O segundo e terceiro grupos consistem de estruturas que fazem indexação em espaços Euclidianos n-dimensionais. A diferença entre os dois grupos está no fato de os MAEP indexarem pontos, enquanto que os MAENP serem capazes de indexar também objetos não-puntuais. A Grid File, a K-D-B-tree, dentre outras estruturas, são exemplos de MAEP, enquanto que a família R-tree (R-tree, R+-tree, Greene R-tree e R*-tree), dentre outras, são exemplos de MAENP. O último grupo é aquele composto por estruturas para indexação em espaços determinados por uma função de distância. Os principais exemplos deste grupo são a M-tree, a Slim-tree e a OMNI-Family. A Slim-tree é o foco deste trabalho.

A proposta deste trabalho é construir uma estrutura de indexação métrica, a ser chamada Fuzzy Slim-tree, que melhora o desempenho dos algoritmos existentes da Slim-tree utilizando conceitos básicos da Lógica Nebulosa.

2. Trabalhos Relacionados

2.1. Indexação Métrica

Na literatura, existe uma quantidade considerável de estruturas para indexação em espaços métricos. Em [4] é proposta uma série de abordagens para indexação e busca de elementos em um espaço métrico de forma recursiva, tal qual uma árvore. A primeira delas determina partições de objetos usando uma função de distância, sendo um dos elementos escolhido como representante deste conjunto. A segunda diz respeito à divisão dos conjuntos originais em conjuntos de tamanho fixo, sendo que para cada conjunto é escolhido um representativo. Na busca, os representativos são utilizados para diminuir o escopo da busca através de um critério de poda obtido a partir da desigualdade triangular. A árvore métrica de Uhlmann [13] e a vp-tree de Yianilos [15] são duas árvores similares que fazem indexação métrica dividindo o espaço em dois a partir de um *vantage point*, de forma que metade dos objetos se encontra dentro do raio do círculo métrico e a outra metade, fora. Ainda em [13], foi proposta a gh-tree (*generalized hyper-plane tree*), que particiona os dados em dois conjuntos e escolhe um representativo para cada um, designando o restante dos objetos ao representativo mais próximo. Para diminuir a quantidade de cálculos de distância, [1] propôs usar o mesmo *vantage point* para todos os nós de um mesmo nível na FQ-tree. Em [3] é proposta a GNAT (*Geometric Near-Neighbor Access Tree*), que possui um número k de *split points* no nível da raiz, sendo cada um deles associado a uma das partições. Para cada *split point*, as distâncias mínima e máxima para cada ponto são guardadas. Em [2] é proposta a mvp-tree (*multi-vantage-tree*), que é uma extensão da vp-tree que utiliza dois *vantage points* por nó.

Todos estes métodos citados são estáticos, de forma que eles não dão suporte a inserções nem exclusões. A M-tree [5] resolve esta deficiência. Ela é uma árvore balanceada, na qual os dados são armazenados nas folhas. Os nós interiores (índice), têm ponteiros que direcionam o acesso aos nós folha. Cada nó possui no máximo `C` objetos e no mínimo `C/2` objetos, sendo um deles chamado representativo, em torno do qual o raio determina a cobertura do nó. Em [7], foi proposta a M^2-tree, que considera mais que uma função de distância, com cada uma delas correspondendo a uma "dimensão" do espaço métrico. Para mensurar o grau de sobreposição que pode existir entre os nós de uma árvore métrica, foi proposta a Slim-tree [11]. Assim, foram criados o *fat-factor* (grau absoluto) e o *bloat-factor* (grau relativo). Um algoritmo chamado Slim-down foi proposto para diminuir a sobreposição entre os nós. Mais recentemente, [10] propõem o conceito OMNI pra construir um método de acesso métrico sobre um outro já existente. A idéia é eleger um conjunto de objetos como referência, cada um determinando um foco, guardando a distância de todos os outros objetos em relação a eles.

2.2. Lógica Nebulosa

A Lógica Nebulosa, definida por Zadeh [16], é a ciência que se preocupa em representar princípios formais do raciocínio aproximado. Seu objetivo é fornecer fundamentos para representar o raciocínio com proposições imprecisas, que não são inteiramente verdadeiras nem falsas, mas um pouco de cada uma, utilizando-se da teoria de conjuntos nebulosos. Assim, para cada elemento de um conjunto nebuloso, é designado um valor dentro do intervalo [0,1], chamado grau de pertinência, que é determinado pela função de pertinência.

A Lógica Nebulosa vem sendo bastante aplicada em técnicas de recuperação de informação. Como exemplo, em [8] e em [14] são efetuadas consultas envolvendo conceitos da

Lógica Nebulosa. Em [6] é definida uma álgebra para busca por similaridade através de operadores que utilizam valores nebulosos.

3. Proposta e Estado Atual

3.1. Slim-tree

A Slim-tree [11] é uma estrutura de indexação métrica, baseada na M-tree, balanceada, que possui nós índice que direcionam o acesso aos nós folha, nos quais os objetos se encontram. Ela foi criada para medir e diminuir o grau de sobreposição entre os nós da árvore métrica. Para fazer a medição foram propostos o *fat-factor* e o *bloat-factor*, que representam o grau absoluto e relativo (à árvore ótima) de sobreposição de uma árvore, respectivamente. Para fazer a diminuição da sobreposição, foi proposto o Slim-down cuja idéia é fazer a reinserção de objetos de forma conveniente a diminuir o raio de cobertura dos nós. A Slim-tree também propôs o algoritmo de inserção por mínima ocupação e o algoritmo MST (*Minimal Spanning Tree*) de divisão de nó e escolha dos representativos.

3.2. Fuzzy Slim-tree

A Fuzzy Slim-tree será uma estrutura de indexação métrica baseada na Slim-tree. Sua estrutura será semelhante à da Slim-tree, diferindo no fato de ter um campo a mais: o Fuzzy(Si), que representará o grau de pertinência da entrada (ou do nó) em relação ao representativo do nó raiz. Além disso, a Fuzzy Slim-tree apresentará algoritmos para inserção, divisão de nó e escolha dos representativos e Slim-down utilizando conceitos da Lógica Nebulosa. As estruturas dos nós índice e folha são mostradas na **Figura 1**.

(a) (b)

Figura 1 – Estruturas dos nós: (a) índice; (b) folha.

A diferença entre a estrutura da Fuzzy Slim-tree e a da Slim-tree é o campo $Fuzzy(S_i)$, presente na primeira e que é dado pela expressão:

$Fuzzy(S_i) = 1$, se S_i é o representativo do nó raiz;

$$Fuzzy(S_i) = 1 - \frac{dist(O_r, S_i)}{R_i}, \text{ para } dist(O_r, S_i) < R;$$

$Fuzzy(S_i) = 0$, caso contrário, onde:

S_i é o objeto, O_r é o representativo do nó raiz e R é o raio de cobertura do nó raiz.

3.3. Grau de Distribuição (GD)

O Grau de Distribuição (GD) será definido como sendo o valor médio de pertinência relativa dos objetos em relação ao representativo do nó raiz e será indicado pela expressão:

$$GD(\text{nó}) = \frac{\sum_{i=1}^{n} Fuzzy(S_i)}{n}, \text{ onde:}$$

S_n é o objeto e n é o número de objetos do nó.

3.4. Inserção

A inserção em uma Slim-tree pode ser feita de três formas. A primeira delas escolhe a sub-árvore de forma aleatória. A segunda (minDist) leva em conta a menor distância do objeto em relação ao representativo. A última (minoccup) seleciona o nó com menor ocupação. Nes-

tes casos, pode haver um empate na escolha, que pode ser resolvido aplicando-se um conjunto de valores Fuzzy $V_I = \{v_{I1}, v_{I2}, \ldots, v_{In}\}$, dado pela expressão:

$v_{Ii} = 1$, se o objeto S_k é igual ao representativo;

$v_{Ii} = 1 - \frac{dist(Rep(S_i), S_k)}{R_i}$, se S_k pertence ao raio de cobertura do nó i e é diferente do representativo.

Através da análise deste valor Fuzzy, pode-se saber o grau de pertinência do objeto em relação a cada um dos nós de um determinado nível. Daí é conveniente escolher o nó em relação ao qual o objeto tiver o maior grau de pertinência, isto é, tiver o maior valor Fuzzy.

3.5. Divisão de Nó e Escolha dos Representativos

Na Slim-tree, a divisão do nó e escolha dos representativos pode ser feita de três formas. A primeira delas escolhe um par de representativos de forma aleatória. A segunda (minMax) calcula todas as possibilidades de pares de nós representativos e escolhe aquele que minimiza o raio de cobertura. A última (MST) gera a árvore geradora mínima do nó a ser dividido, elimina um dos maiores arcos e escolhe o objeto de cada árvore gerada da divisão que minimiza o raio de cobertura. Neste último caso, pode haver um empate na escolha, que pode ser resolvido calculando-se o Grau de Distribuição para cada uma das possibilidades. Aquela configuração que tiver a menor média dos Graus de Distribuição dos nós deve ser a escolhida.

3.6. Fuzzy Slim-down

O Slim-down é um algoritmo que propõe diminuir o grau de sobreposição entre nós em uma árvore métrica. Para cada nó em um dado nível da árvore, verificar se o objeto mais distante está sob o raio de cobertura de outro nó. Em caso positivo, eliminar este objeto e reinseri-lo no outro nó sob o raio do qual ele se encontrava. Repetir o processo. Pode haver um empate na escolha do nó no qual o objeto deve ser inserido. Para resolver este impasse, é proposto o algoritmo Fuzzy Slim-down, que utiliza um conjunto de valores Fuzzy da mesma forma que na inserção. De forma semelhante, deve-se escolher o nó em relação ao qual o objeto tiver o maior grau de pertinência.

3.7. Consulta Fuzzy Range Query

Tomando como ponto de partida a estrutura da **seção 3.2**, na qual o campo `Fuzzy(Si)` corresponde à posição relativa da entrada em relação ao representativo do nó raiz, é proposto o algoritmo de consulta Fuzzy Range Query em domínios métricos que diminui o número de cálculos de distância como um todo. Dados o representativo do nó raiz O_r de uma árvore métrica e um conjunto de nós de um certo nível da árvore, ao se realizar uma consulta sobre um objeto `Q` de consulta e `r(Q)` o raio da consulta, pode-se utilizar as distâncias $d(Q,Or) \pm r(Q)$ como fator de "poda", evitando que cálculos de distância sejam feitos.

3.8. Consulta Fuzzy k-Nearest Neighbor Query

Tomando como ponto de partida a estrutura da **seção 3.2**, na qual o campo `Fuzzy(Si)` corresponde à posição relativa da entrada em relação ao representativo do nó raiz, é proposto o algoritmo de consulta Fuzzy k-Nearest Neighbor Query em domínios métricos que diminui o número de cálculos de distância efetuados. Dados o representativo do nó raiz O_r de uma árvore métrica e um conjunto de nós de um certo nível da árvore, ao se realizar uma consulta sobre um objeto `Q` de consulta, é utilizada uma lista de prioridades de acordo com a distância ascendente do nó/objeto em relação a O_r, de forma que o primeiro nó a ser consultado é o mais próximo e o último é o mais distante.

4. Resultados Esperados

Com este trabalho, pretende-se utilizar conceitos de Lógica Nebulosa nos algoritmos de inserção, divisão de nó e escolha dos representativos e Slim-down para melhorar a distribuição dos objetos e diminuir o grau de sobreposição de uma árvore métrica. Por conseqüência, pretende-se melhorar o desempenho da árvore métrica Slim-tree, que está relacionado ao número de páginas acessadas, à quantidade de cálculos de distância e à utilização do espaço de armazenamento. Todos esses parâmetros de melhoria indicados (tópicos 3.1 a 3.8) devem ser implementados para melhorar o estado da árvore com relação à realização das consultas.

Referências Bibliográficas

[1] Baeza-Yates, R. / Cunto, W. / Mamber, U. / Wu, S.: "Proximity Matching Using Fixed-Queries Trees", 5th Symposium on Combinatorial Pattern Matching, p. 198-212, 1994.
[2] Bozkaya, T. / Ozsoyoglu, M.: "Distance-Based Indexing for High Dimensional Metric Spaces", ACM International Conference on Management of Data, p. 357-368, 1997.
[3] Brin, S.: "Near Neighbor Search in Large Metric Spaces", 21st International Conference on Very Large Data Bases (VLDB), p. 574-584, 1995.
[4] Burkhard, W. A. / Keller, R. M.: "Some Approaches to Best-Match File Searching", Communications of the ACM, v. 16, no. 4, p. 230-236, 1973.
[5] Ciaccia, P. / Patella, M. / Zezula, P.: "M-tree: An Efficient Access Method for Similarity Search in Metric Spaces", 23th International Conference on Very Large Data Bases (VLDB), p. 426-435, 1997.
[6] Ciaccia, P. / Montesi, D. / Penzo, W. / Trombetta, A.: "Fuzzy Query Languages for Multimedia Data", Design and Management of Multimedia Information Systems: Oportunities and Challenges, Idea Group Publishing, 2001.
[7] Ciaccia, P. / Patella, M.: "The M^2-Tree: Processing Complex Multi-Feature Queries with Just One Index", 1st DELOS Network of Excellence Workshop on information Seeking, Searching and Querying in Digital libraries, 2000.
[8] Fagin, R.: "Fuzzy Queries in Multimedia Database Systems", 17th ACM Symposium on Principles of Database Systems (PODS'98), p. 1-10, 1998.
[9] Gaede, V. / Günther, O.: "Multidimensional Access Methods", ACM Computing Surveys, 1997.
[10] Santos Filho, R. F. / Traina, A. / Traina Jr., Caetano / Faloutsos, C.: "Similarity Search Without tears: The OMNI Family of All-purpose Access Methods", 17th Conference on Data Engineering (ICDE'2001), 2001.
[11] Traina Jr., C. / Traina, A. / Seeger, B. / Faloutsos, C.: "Slim-trees: High Performance Metric Trees Minimizing Overlap Between Nodes", 8th International Conference Extending Database Technology (EDBT'00), p. 41-65, 2000.
[12] Traina Jr., C. / Traina, A. / Faloutsos, C. / Seeger, B.: "Fast Indexing and Visualization of Metric Data Sets using Slim-Trees", IEEE Transactions on Knowledge and Data Engineering, v. 14, no. 2, p. 244-260, 2002.
[13] Uhlmann, J. K.: "Satisfying general Proximity/Similarity Queries with Metric Trees", Information Processing Letters (IPL), v. 40, no. 4, p. 175-179, 1991.
[14] Vieira, M. T. P. / Biajiz, M. / Borges Jr., Sérgio Ricardo / Teixeira, E. C. / dos Santos, F. G. / Figueiredo, J. M.: "Content-Based Fuzzy Search in a Multimedia Web Database", Intelligent Exploration of the Web, "Studies in Fuzziness and Soft Computing" Series, 2002.
[15] Yianilos, P. N.: "Data Structures and Algorithms for Nearest Neighbor Search in General Metric Spaces", ACM SIGACT-SIAM Symposium on Discrete Algorithms, 1993.
[16] Zadeh, L. A.: "Fuzzy Sets", Information and Control, v. 8, no. 3, p. 338-353, 1965.