Estudos para o desenvolvimento de uma biblioteca de implementação de árvores R com aplicação a mineração de dados

Miguel dos Santos Pereira

Faculdade de Computação
Universidade Federal de Uberlândia

4 de junho de 2014

Objetivo

- Estudar R-tree;
- Aplicação em bancos de dados multidimensionais;
- Implementação de uma biblioteca;
- Resolver problemas de classificação;

Contexto de Surgimento da R-tree

Contexto

- B-trees não eram suficientes;
- Novas áreas;
- Novos tipos de dados;

Surgimento da R-tree

Surgimento

- Surgiu uma estrutura chamada R-tree;
- Focada em suportar dados geométricos/espaciais;
- Utilização multi-dimensionais;
- Novas modificações foram sugeridas;

Usos da R-tree

Atualmente

- Sistemas espaciais e geométricos estão em alta;
- Sistemas espaço-temporais e multimídia estão sendo estudados;
- Novos tipos de dados devem considerar usar R-trees;

Ideia da R-tree

Ideia Chave

- Agrupar objetos próximos;
- Representa-los em um retângulo de menor área possível;
- Se um objeto não intercepta um retângulo, também não intercepta nenhum de seus filhos;
- No nível da folha, cada retângulo descreve um único objeto;
- Ao se aproximar da raiz, pode ser visto um "conjunto de dados" de forma grosseira;

Ideia da R-tree

Balanceamento

- Similar a B-trees, R-trees são balanceadas, paginadas e utilizam armazenamento em disco;
- Cada página pode conter um número máximo (M) de entradas;
- Isto também garante um preenchimento mínimo;

Ideia da R-tree

Obstáculos

- Construir uma árvore balanceada;
- Retângulos não cubram muito espaço vazio;
- Retângulos não se sobreponham muito;

Operações

Objetivo

- Inserir o nodo em uma subárvore que requer menor alargamento de seu retângulo;
- Uma vez que a página está cheia, os dados são separados em dois conjuntos que devem cobrir uma área mínima cada;

Características

Uma página está cheia quando:

- Em uma árvore R de ordem (m, M), o nodo tiver M entradas;
- Sendo $m < M/2$ o número mínimo de entradas permitidas;

Glossário

MBR = Minimum Bounding Rectangle (Retângulo de Limite Mínimo).
$E.mbr = MBR$ da entrada E.

Inserção simplificada

```
1  Inserção(TipoDeEntrada E, TipoDoNodo RN) {
2    Faz travessia da Raiz RN até Nodo L cujo o MBR de menor alargamento para
         cobrir E.mbr
3
4    if (Nodo L está cheio)
5      divide em dois de tamanho M
6
7    Propaga alterações de área e número de elementos para cima da árvore
8
9    if (raiz ficar cheia)
10     divide raiz e aumenta altura
11 }
```

Inserção detalhada

1. Faça a travessia da Raiz RN para inserir a entrada E na folha apropriada.
   - A cada nível, siga subnodo com MBR que precisará do menor alargamento para cobrir E
   - No caso de empate, escolha o Nodo com MBR de menor área.
2. **Se** a folha L não está cheia, insira E a L
3. **Se** L está cheia
   - Seja E o conjunto de todas as entradas de L mais a entrada E
   - Selecione como sementes duas entradas $e_1, e_2 \in E$, onde a distância entre e_1 e e_2 seja a maior entre pares em E sendo $e_1 \in L_1$ e $e_2 \in L_2$, potenciais novos nodos
   - Atribua $e \in E$ restantes ao nodo que requerer o menor alargamento para cubrir e
     - No caso de empate, atribua a entrada no Nodo que tenha menor MBR.
     - **Se** novo empate, atribua a entrada no Nodo que tenha menor número de entradas.
     - **Se** durante as atribuições, sobrarem k entradas para serem atribuídas e um nodo tem $m - k$ entradas, atribua o restante das entradas para o outro nodo
4. Atualize as $MBRs$ dos nodos que estão entre a Raiz e o L (ou L_1 e L_2)
5. Faça separações nos níveis superiores se necessário.
6. Caso raiz precise ser dividida, crie uma nova Raiz e aumente a altura da árvore em 1.

KNN

Objetivo

Fazer identificação de classes.

Método

Usar classes mais frequentes dos vizinhos

Exemplo: Identificar a classe do Carro

Atributos de um Carro:

- Nº Rodas, Nº Portas, Potência, Nº Passageiros, Preço

Classes pré-definidas:

- Van, SPORT, Sedan, HATCH

Etapas de trabalho

- Estudos sobre árvores de busca tradicionais
- Estudos sobre árvore R
- Implementação da árvore R (em Java)
- Testes de desempenho com diferentes versões da construção da árvore R (linear, quadrático)
- Implementação algoritmo k-nearest neighbors KNN (aprendizado baseado em instâncias)
- Aplicação do algoritmo KNN com árvore R para conjuntos de dados de diferentes dimensionalidades

Referências

Y.Manopoulos, A.Nanopoulos, A.N.Papadoulos and Y.Theodoridis: "R-Trees: Theory and Applications", pp.24-31, Grécia, 2006.
Wikipedia contributors. "R-tree.", Wikipedia, The Free Encyclopedia, 2014