Consultas por Similaridade em Domínios de Dados Complexos

Guilherme Felipe Zabot

Orientador: Profº. Dr. Caetano Traina Jr.

Roteiro

- Objetivo
- Introdução
- Dados Complexos
- Representação de Dados Complexos
- Extração de Características
- Medidas de Similaridade
- Espaço Métrico
- Funções de Distância
  - Minkowski
  - Métricas L_p
- Consultas por Similaridade
  - *Range query*
  - *K-nearest neighbor query*
- Métodos de Acesso Métrico
- Técnica Omni

Objetivo da apresentação

O que é Recuperação de Dados por Similaridade?

É recuperar os elementos de uma coleção de dados, que são similares entre si ou aqueles que se conhecem.

Objetivo desta apresentação:

Apresentar conceitos e técnicas necessárias para responder consultas por similaridade em domínios de **dados complexos**.

Introdução

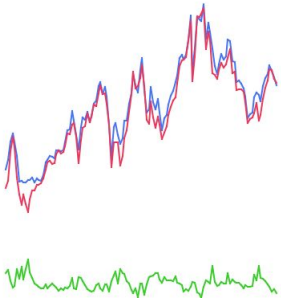
- Nos últimos anos houve um aumento na quantidade e complexidade dos dados.

Introdução

- Nos últimos anos houve um aumento na quantidade e complexidade dos dados.
- Não estruturados, requerem um pré-processamento para manipulá-los.

Introdução

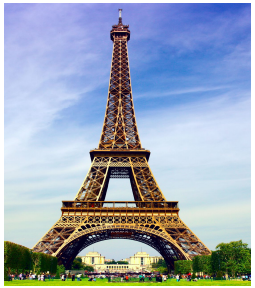
- Nos últimos anos houve um aumento na quantidade e complexidade dos dados.
- Não estruturados, requerem um pré-processamento para manipulá-los.
- Exemplo de dados complexos:



Séries temporais

Músicas

Vídeos



Imagens

Dados Complexos

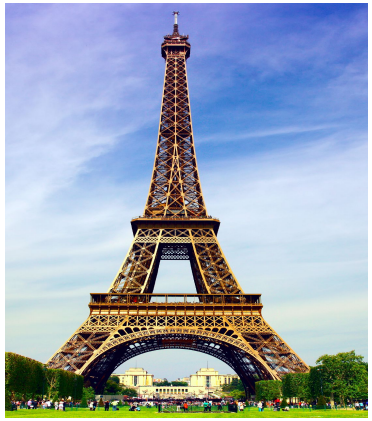
- Dados complexos não são estruturados, como o caso de idade, anos, datas, etc.

Dados Complexos

- Dados complexos não são estruturados, como o caso de idade, anos, datas, etc.
- Desenvolvimento de técnicas para recuperar e manipular esses tipos de dados.

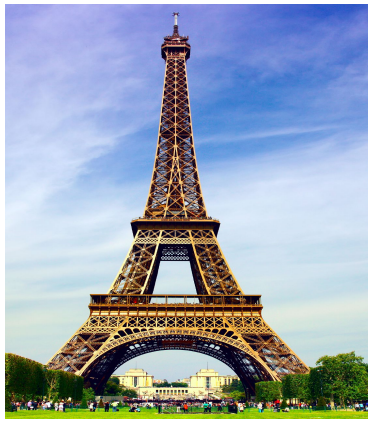
Dados Complexos

- Dados complexos não são estruturados, como o caso de idade, anos, datas, etc.
- Desenvolvimento de técnicas para recuperar e manipular esses tipos de dados.
- Consultas por similaridade têm sido utilizadas como forma de manipular dados complexos.



Dados Complexos

- Dados complexos não são estruturados, como o caso de idade, anos, datas, etc.
- Desenvolvimento de técnicas para recuperar e manipular esses tipos de dados.
- Consultas por similaridade têm sido utilizadas como forma de manipular dados complexos.



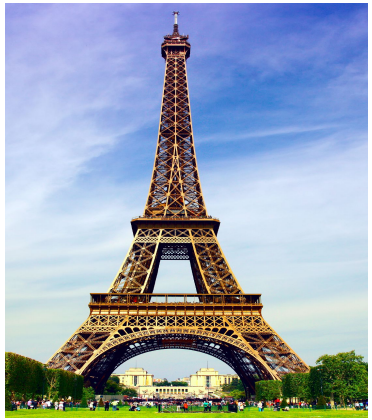
?

?

Dado essas três imagens, como compará-las?

Dados Complexos

- As consultas baseadas em relação de identidade (RI) ou em relação de ordem (RO), úteis para dados escalares, não são adequadas para dados complexos.

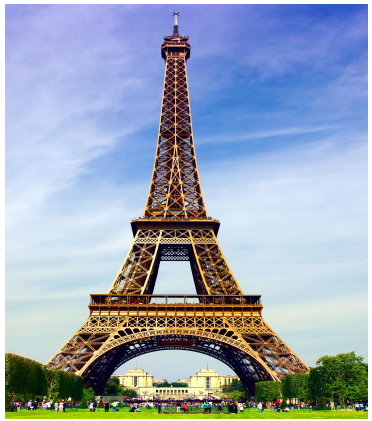


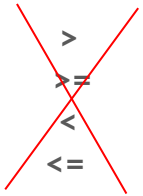
> >= < <=

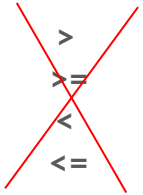
> >= < <=

Dados Complexos

- As consultas baseadas em relação de identidade (RI) ou em relação de ordem (RO), úteis para dados escalares, não são adequadas para dados complexos.

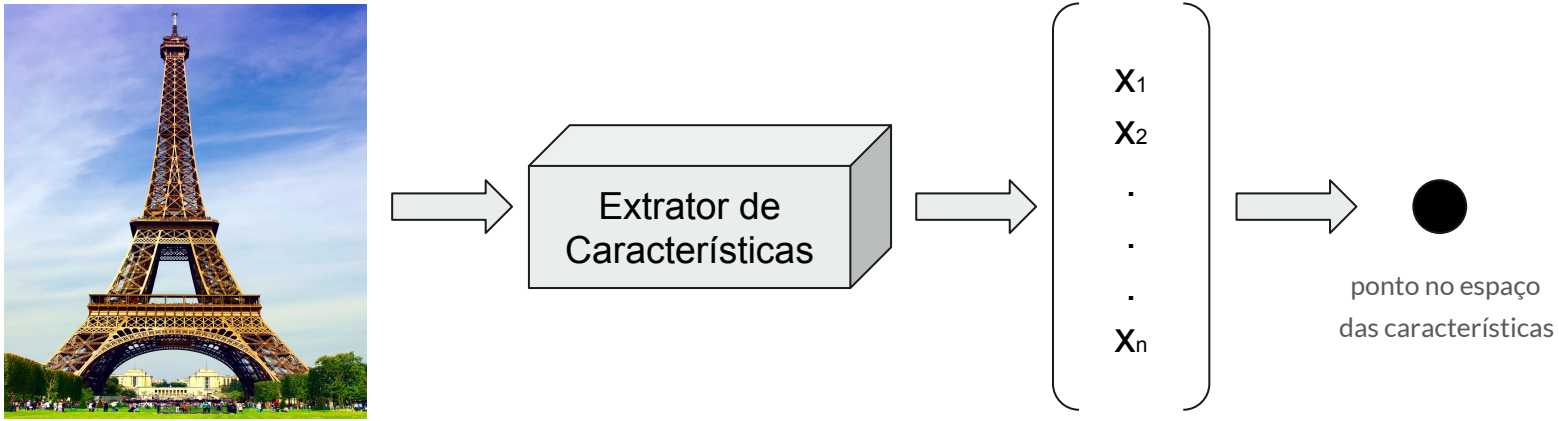




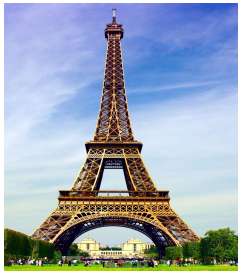


Representação de Dados Complexos

- Representação dos dados: Vetor de características (feature vector).



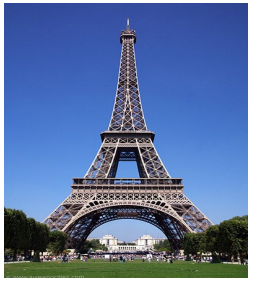
Representação de Dados Complexos



$$\begin{bmatrix} x_1 \\ x_2 \\ . \\ . \\ . \\ xn \end{bmatrix}$$

Vetor de características

Ponto no espaço das características



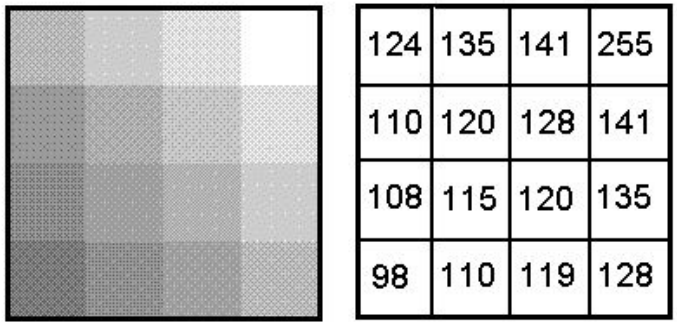
Extração de Características

- Para serem computacionalmente utilizadas em consultas, imagens devem ser primeiro processadas por meio de extratores de características para a criação de vetores de características.

Extração de Características

- Para serem computacionalmente utilizadas em consultas, imagens devem ser primeiro processadas por meio de extratores de características para a criação de vetores de características.

- Uma imagem é representada como uma matriz bidimensional *m* x *n* de pixels, onde m e n são as dimensões da imagem, e cada pixel possui valores inteiros que dependem do tipo da imagem.



124	135	141	255
110	120	128	141
108	115	120	135
98	110	119	128

Ilustração de uma imagem de tamanho 4 x 4 pixels.
Fonte: http://www.informaticamedica.org.br/informaticamedica/n0106/imagens.htm

Extração de Características

- Nas áreas de processamento de imagens e visão computacional são estudados possíveis descritores de imagens, os quais representam as imagens com base principalmente nas propriedades de cor, textura e forma.

Extração de Características

- Nas áreas de processamento de imagens e visão computacional são estudados possíveis descritores de imagens, os quais representam as imagens com base principalmente nas propriedades de cor, textura e forma.

- Os algoritmos que implementam esses descritores, ou seja, realizam o processo de extração de características, são chamados de extratores.

Extração de Características

- Nas áreas de processamento de imagens e visão computacional são estudados possíveis descritores de imagens, os quais representam as imagens com base principalmente nas propriedades de cor, textura e forma.

- Os algoritmos que implementam esses descritores, ou seja, realizam o processo de extração de características, são chamados de extratores.

- **Cor:**
  - Dentre as técnicas existentes, o histograma de cor é o mais simples e o mais utilizado.

Extração de Características

- Nas áreas de processamento de imagens e visão computacional são estudados possíveis descritores de imagens, os quais representam as imagens com base principalmente nas propriedades de cor, textura e forma.

- Os algoritmos que implementam esses descritores, ou seja, realizam o processo de extração de características, são chamados de extratores.

- **Cor:**
  - Dentre as técnicas existentes, o histograma de cor é o mais simples e o mais utilizado.

- **Textura:**
  - É possível identificar padrões repetitivos sobre a superfície da imagem, medindo propriedades tais como aspereza, suavidade e regularidade

Extração de Características

- Nas áreas de processamento de imagens e visão computacional são estudados possíveis descritores de imagens, os quais representam as imagens com base principalmente nas propriedades de cor, textura e forma.

- Os algoritmos que implementam esses descritores, ou seja, realizam o processo de extração de características, são chamados de extratores.

- **Cor:**
  - Dentre as técnicas existentes, o histograma de cor é o mais simples e o mais utilizado.

- **Textura:**
  - É possível identificar padrões repetitivos sobre a superfície da imagem, medindo propriedades tais como aspereza, suavidade e regularidade

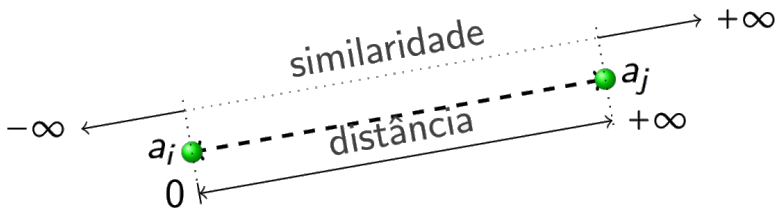
- **Forma**
  - São categorizados em descritores baseados em contorno e descritores baseados em região. Enquanto no primeiro apenas a borda dos objetos é explorada, no segundo todos os pixels dentro de uma região são levados em conta.

Medidas de Similaridade

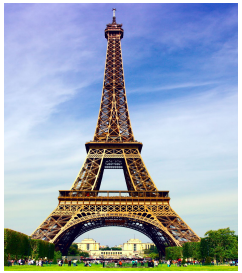
- Para realizar as consultas utilizando essas estruturas é necessário definir uma função que meça o grau de semelhança entre esses vetores.

Medidas de Similaridade

- Para realizar as consultas utilizando essas estruturas é necessário definir uma função que calcule o grau de semelhança entre esses vetores.
    - Função de Similaridade:
        - O resultado da função é um valor numérico que é **maior** para elementos mais similares;
    - Função de Distância:
        - O resultado da função é um valor numérico não negativo que é **menor** para elementos mais similares.



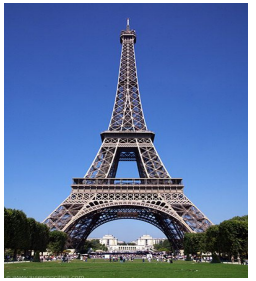
Medidas de Similaridade



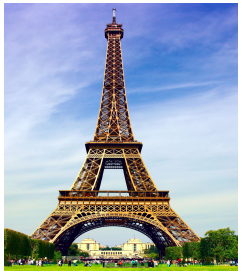
x_1
x_2
.
.
.
x_n

Vetor de características

Ponto no espaço das características



Medidas de Similaridade



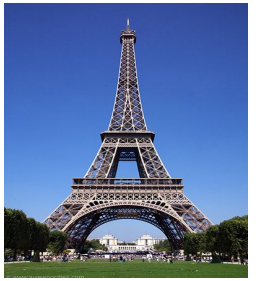
x_1
x_2
.
.
.
xn

Vetor de características

Menor distância

Ponto no espaço das características

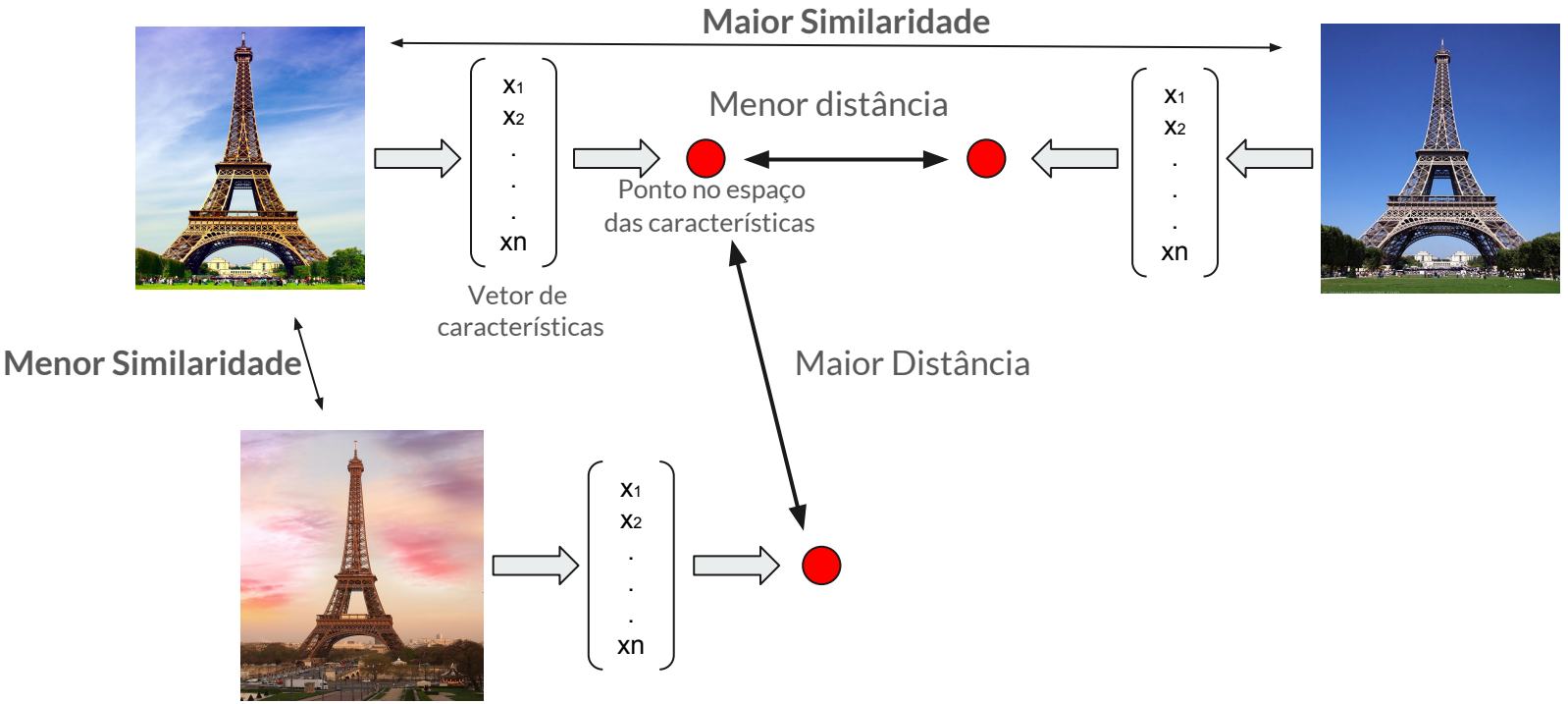
x_1
x_2
.
.
.
xn



Maior Distância

x_1
x_2
.
.
.
xn

Medidas de Similaridade

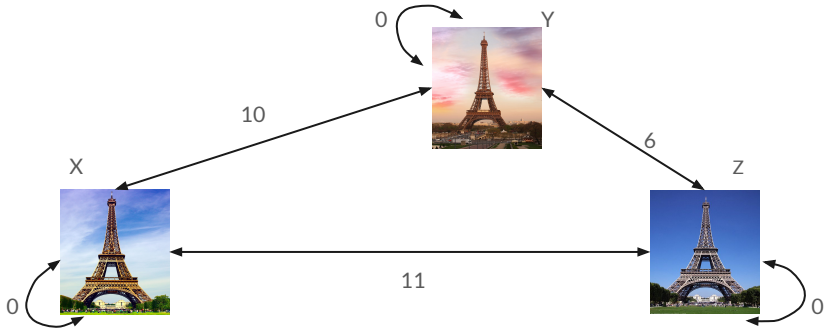


Espaço Métrico

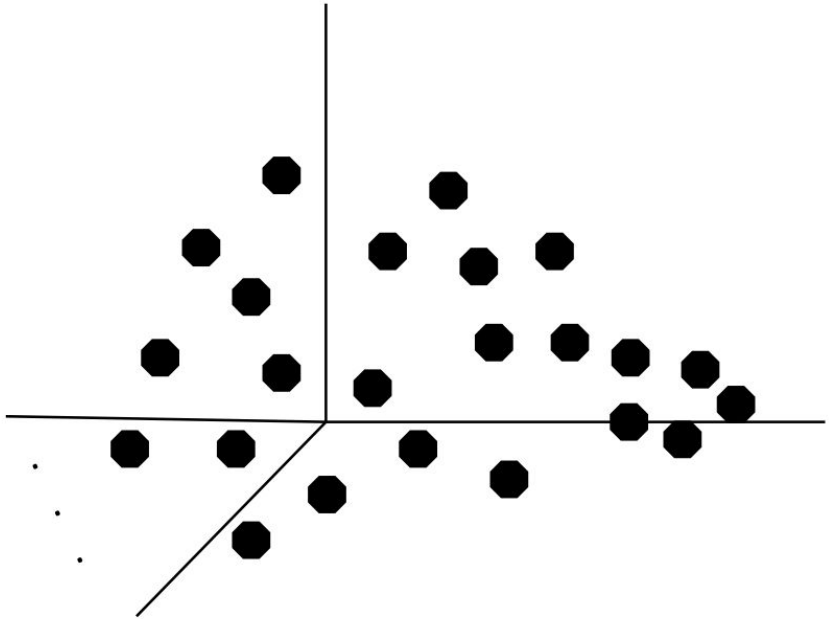
- Um espaço métrico é um par $M = D, d(\ )$, onde D é um conjunto de elementos (as chaves de indexação) e $d(\ )$ uma função de distância (ou métrica).

Espaço Métrico

- Um espaço métrico é um par $M = D, d(\ )$, onde D é um conjunto de elementos (as chaves de indexação) e $d(\ )$ uma função de distância (ou métrica).
    - Identidade ($\forall x \in U, d(x,x) = 0$);
    - Simetria: $d(x, y) = d(y, x)$;
    - Não-negatividade $0 \leq d(x, y) < \infty$;
    - Desigualdade triangular $d(x, y) \leq d(x, z) + d(z, y)$;



Espaço Métrico



Espaço Métrico

- Dessa forma, um vetor de características representado neste modelo pode ser visto como um ponto em um espaço n-dimensional, onde n é o tamanho do vetor.

- Com isso , é possível representar certos tipos de objetos, como impressões digitais, pois o número de características que podem ser extraídas variam entre as pessoas.

Funções de Distância

- As funções de distância expressam numericamente o grau de dissimilaridade entre dois objetos.

Funções de Distância

- As funções de distância expressam numericamente o grau de dissimilaridade entre dois objetos.
- Dessa forma, a similaridade é inversamente proporcional ao resultado dessa função, isto é, quanto menor a distância obtida, maior a similaridade.

Funções de Distância

- As funções de distância expressam numericamente o grau de dissimilaridade entre dois objetos.
- Dessa forma, a similaridade é inversamente proporcional ao resultado dessa função, isto é, quanto menor a distância obtida, maior a similaridade.
- Apesar de existirem diversas funções de distância para vários tipos de aplicações presentes na literatura, as distâncias da família Minkowski são as mais utilizadas.

Funções de Distância

- As funções de distância expressam numericamente o grau de dissimilaridade entre dois objetos.
- Dessa forma, a similaridade é inversamente proporcional ao resultado dessa função, isto é, quanto menor a distância obtida, maior a similaridade.
- Apesar de existirem diversas funções de distância para vários tipos de aplicações presentes na literatura, as distâncias da família Minkowski são as mais utilizadas.
- Conhecidas também como norma Lp, essa família de distâncias é válida para espaços vetoriais, nos quais os objetos são definidos por meio de tuplas $\{x_1, x_2, ..., x_n\}$, onde *n* é o tamanho do vetor (quantidade de características).

$$L_p((x_1, ..., x_n), (y_1, ..., y_n)) = \sqrt[p]{\sum_{i=1}^{n} |x_i - y_i|^p}$$

Definição da equação da família de distâncias

Funções de Distância Minkowski ou métricas L_p

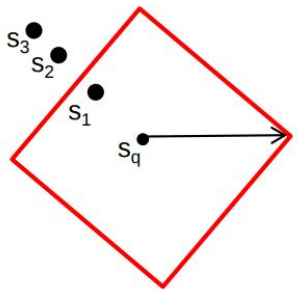
- Há alguns casos particulares desses grupos de funções, que podem ser obtidos à partir de variações do parâmetro p.
    - Para p=1, tem-se a distância L_1, também chamada de City Block ou distância *Manhattan*.
    - Para p=2, tem-se a distância L_2, conhecida como distância Euclidiana.
    - Ao se calcular o limite da função L_p quando p tende ao infinito, obtém-se a função L_∞, chamada de Infinity ou *Chebychev*.

Funções de Distância Minkowski ou métricas L_p

- Há alguns casos particulares desses grupos de funções, que podem ser obtidos à partir de variações do parâmetro p.
    - Para p=1, tem-se a distância L_1, também chamada de City Block ou distância *Manhattan*.
    - Para p=2, tem-se a distância L_2, conhecida como distância Euclidiana.
    - Ao se calcular o limite da função L_p quando p tende ao infinito, obtém-se a função L_∞, chamada de Infinity ou *Chebychev*.
- Cada função de distância da família Minkowski gera, ao redor de um elemento de consulta *q*, uma forma geométrica diferente formada pelos pontos equidistantes a um dado radio de abrangência *rq*.

Funções de Distância Minkowski ou métricas L_p

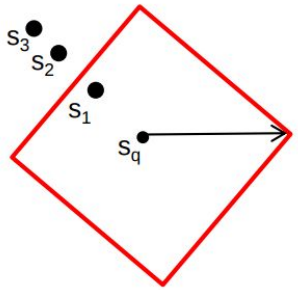
Considerando um espaço bidimensional



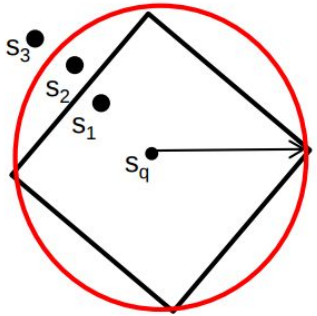
L_1 – distância Manhattan

Funções de Distância Minkowski ou métricas L_p

Considerando um espaço bidimensional



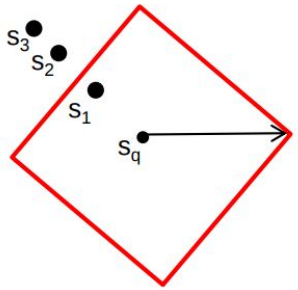
L_1 – distância Manhattan



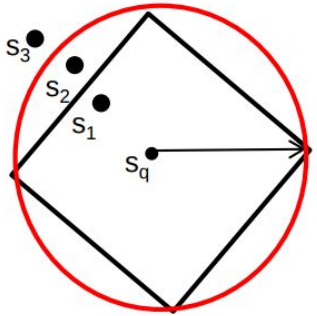
L_2 - distância Euclidiana

Funções de Distância Minkowski ou métricas L_p

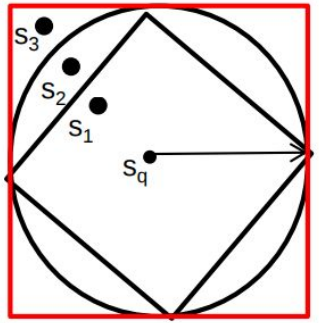
Considerando um espaço bidimensional



L_1 – distância Manhattan



L_2 - distância Euclidiana



L_∞ - distância Máxima

Consultas por Similaridade

- Em espaços métricos, os dois tipos comumente utilizados na literatura são:
  - *Range query*;
  - *K-nearest neighbor query*.

Consultas por Similaridade

- Em espaços métricos, os dois tipos comumente utilizados na literatura são:
  - *Range query*;
  - *K-nearest neighbor query*.

- *Range query* R(q, r): Consulta por abrangência, a qual retorna todos os elementos que estão dentro da distância de r a q.

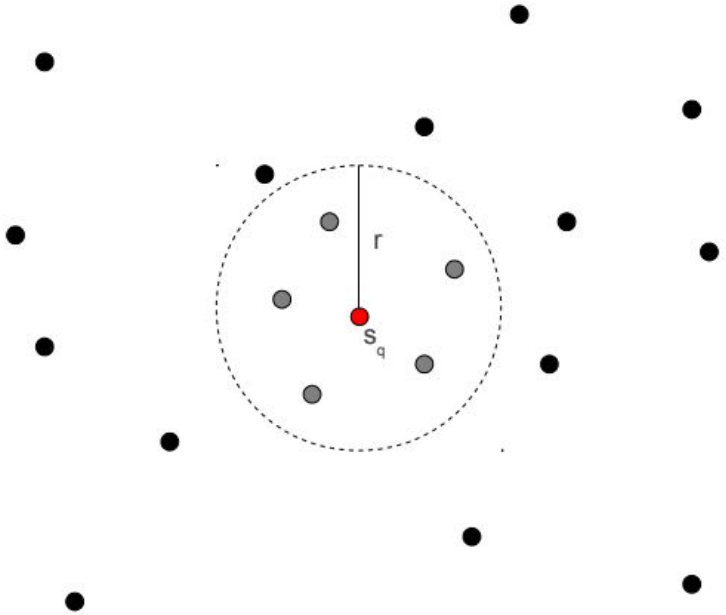
Consultas por Similaridade

- Em espaços métricos, os dois tipos comumente utilizados na literatura são:
  - *Range query*;
  - *K-nearest neighbor query*.

- *Range query* R(q, r): Consulta por abrangência, a qual retorna todos os elementos que estão dentro da distância de r a q.

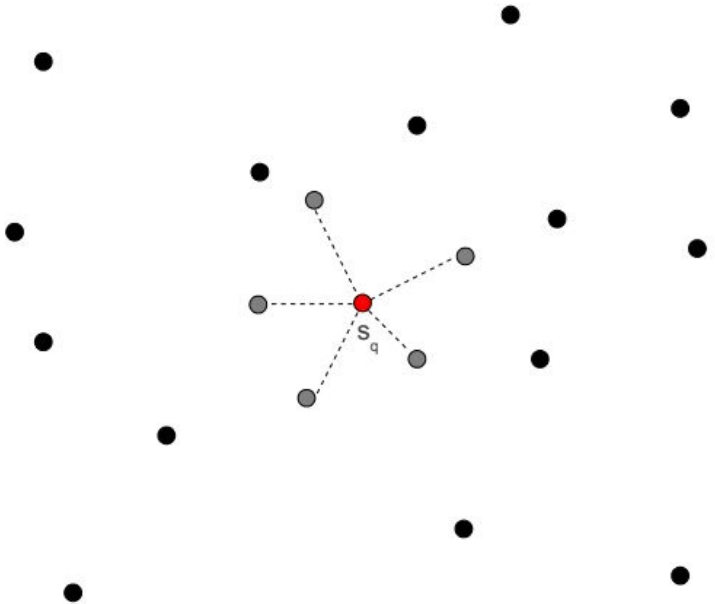
- *K-nearest neighbor query* - Knn(q): Busca pelos vizinhos mais próximos, à qual retorna os k elementos que estão mais perto de q.

Consultas por Similaridade



Consulta por abrangência.

Consultas por Similaridade



Consulta pelos 5 vizinhos mais próximos.

Métodos de Acesso Métrico

- Consultas por similaridade entre objetos complexos são geralmente muito caras.

Métodos de Acesso Métrico

- Consultas por similaridade entre objetos complexos são geralmente muito caras.
- Portanto, ter o apoio de estruturas de indexação para reduzir o número de operações de comparação é muito importante.

Métodos de Acesso Métrico

- Consultas por similaridade entre objetos complexos são geralmente muito caras.

- Portanto, ter o apoio de estruturas de indexação para reduzir o número de operações de comparação é muito importante.
- Apenas as propriedades dos espaços métricos podem ser usadas!
  - Identidade
  - Simetria
  - Não-negatividade
  - Desigualdade triangular

Métodos de Acesso Métrico

- Consultas por similaridade entre objetos complexos são geralmente muito caras.

- Portanto, ter o apoio de estruturas de indexação para reduzir o número de operações de comparação é muito importante.
- Apenas as propriedades dos espaços métricos podem ser usadas!
  - Identidade
  - Simetria
  - Não-negatividade
  - Desigualdade triangular

- Para isso, empregam-se Métodos de Acesso Métricos - MAM.

Métodos de Acesso Métrico

- Consultas por similaridade entre objetos complexos são geralmente muito caras.

- Portanto, ter o apoio de estruturas de indexação para reduzir o número de operações de comparação é muito importante.
- Apenas as propriedades dos espaços métricos podem ser usadas!
  - Identidade
  - Simetria
  - Não-negatividade
  - Desigualdade triangular

- Para isso, empregam-se Métodos de Acesso Métricos - MAM.

- MAM's são índices ou técnicas de indexação projetados para atuarem como um caminho otimizado aos dados, evitando a análise exaustiva durante a recuperação dos elementos que satisfazem a uma consulta.

Métodos de Acesso Métrico

Existem diversas Estruturas de Indexação Métricas:

- Baseadas em memória:
  - gh-tree,
  - vp-tree,
  - onion-tree,
  - ...

- Baseadas em disco:
  - Estáticas:
    - vp-tree e variações,
    - mvp-tree,
    - ...
  - E dinâmicas:
    - M-tree e variações,
    - Slim-tree, e variações
    - ...

A Família Omni para Métodos de Acesso Métrico

- Com o auxílio das MAM's, as buscas são restringidas as porções do conjunto de dados nas quais os objetos armazenados têm maior probabilidade de serem similares a um elemento de consulta.

A Família Omni para Métodos de Acesso Métrico

- Com o auxílio das MAM's, as buscas são restringidas as porções do conjunto de dados nas quais os objetos armazenados têm maior probabilidade de serem similares a um elemento de consulta.

- A técnica Omni consiste de um mecanismo de filtragem baseado no uso de representantes globais do conjunto de dados (**elementos representativos**).

A Família Omni para Métodos de Acesso Métrico

- Com o auxílio das MAM's, as buscas são restringidas as porções do conjunto de dados nas quais os objetos armazenados têm maior probabilidade de serem similares a um elemento de consulta.

- A técnica Omni consiste de um mecanismo de filtragem baseado no uso de representantes globais do conjunto de dados (**elementos representativos**).

- Para cada consulta, esses elementos especificam uma região do espaço métrico na qual garantidamente os objetos mais similares ao objeto de consulta residem.

A Família Omni para Métodos de Acesso Métrico

- Com o auxílio das MAM's, as buscas são restringidas as porções do conjunto de dados nas quais os objetos armazenados têm maior probabilidade de serem similares a um elemento de consulta.

- A técnica Omni consiste de um mecanismo de filtragem baseado no uso de representantes globais do conjunto de dados (**elementos representativos**).

- Para cada consulta, esses elementos especificam uma região do espaço métrico na qual garantidamente os objetos mais similares ao objeto de consulta residem.

- Dessa forma, uma vez que a poda no espaço métrico é realizada, os objetos pertencentes a região encontrada são considerados candidatos à resposta da consulta.

A Família Omni para Métodos de Acesso Métrico

Principal ideia da técnica Omni consiste em selecionar um conjunto de *h* elementos como representantes globais e armazenar, para os demais objetos da base de dados, as respectivas distâncias a cada um dos representantes escolhidos.

A Família Omni para Métodos de Acesso Métrico

- Cada elemento tem suas coordenadas Omni armazenadas numa estrutura de dados. Por exemplo, em um arquivo sequencial:

C_{11}	C_{12}	...	C_{1i}
C_{21}	C_{22}	...	C_{2i}
C_{31}	C_{32}	...	C_{3i}
C_{41}	C_{42}	...	C_{4i}

- Durante as buscas, usa-se a propriedade da desigualdade triangular para podar cálculos de distâncias

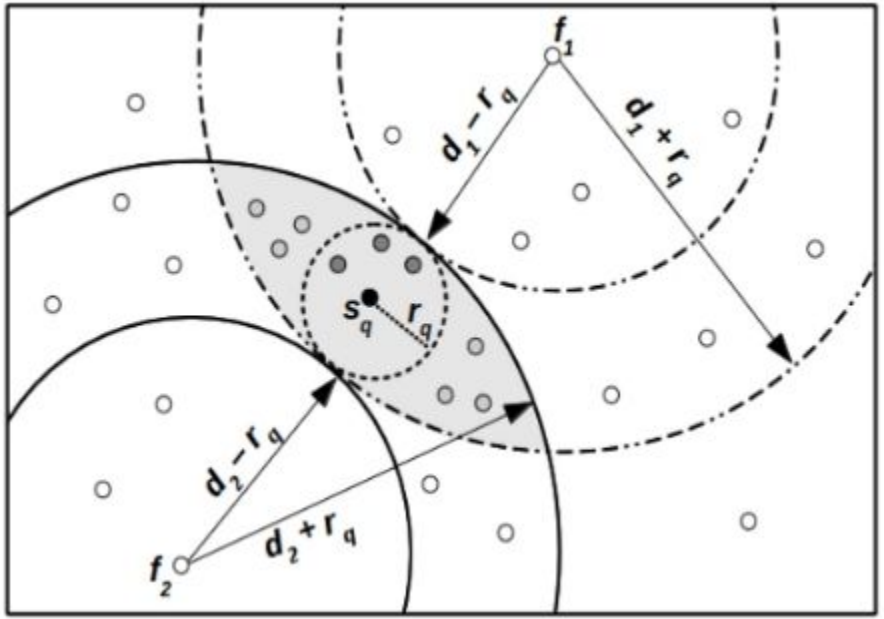
A Família Omni para Métodos de Acesso Métrico

- Consultas por abrangência são executadas em dois passos: **Filtragem e Refinamento.**

A Família Omni para Métodos de Acesso Métrico

- Consultas por abrangência são executadas em dois passos:
  - **Filtragem :**
    - Consiste na seleção de objetos candidatos à resposta com o auxílio dos representantes globais.
    - Esse processo é realizado por meio da especificação de uma ***mbOr*** (*minimum-bounding-Omni-region*).

A Família Omni para Métodos de Acesso Métrico



Definição de uma ***mbOr*** (área hachurada) para uma consulta ao elemento s_q, Com raio r_q e dois elementos representativos, f_1 e f_2.

A Família Omni para Métodos de Acesso Métrico

- Consultas por abrangência são executadas em dois passos:
  - **Refinamento:**

    - Consiste no cálculo efetivo das distâncias entre o elemento de consulta e os objetos contidos na *mbOr* para determinar a resposta da busca.

    - Nota-se que a distância real entre os candidatos e o objeto de consulta é realizada apenas nessa etapa, quando a maioria dos objetos da base de dados já foi descartada.

A Família Omni para Métodos de Acesso Métrico

- Dentre os fatores que influenciam na delimitação de uma boa *mbOr* estão a quantidade de representantes globais e o posicionamento deles no espaço métrico.

A Família Omni para Métodos de Acesso Métrico

- Dentre os fatores que influenciam na delimitação de uma boa *mbOr* estão a quantidade de representantes globais e o posicionamento deles no espaço métrico.

- Um conjunto de elementos representantes é considerado adequado quando ele reduz a *mbOr* o máximo possível, diminuindo a quantidade de distâncias a serem calculadas na etapa de refinamento.

A Família Omni para Métodos de Acesso Métrico

- Dentre os fatores que influenciam na delimitação de uma boa *mbOr* estão a quantidade de representantes globais e o posicionamento deles no espaço métrico.

- Um conjunto de elementos representantes é considerado adequado quando ele reduz a *mbOr* o máximo possível, diminuindo a quantidade de distâncias a serem calculadas na etapa de refinamento.

- Na proposta da técnica Omni, foi mostrado que o número de representantes pode ser obtido de acordo com a dimensionalidade intrínseca do conjunto de dados, que consiste da quantidade mínima de atributos necessária para representar os objetos no conjunto.

A Família Omni para Métodos de Acesso Métrico

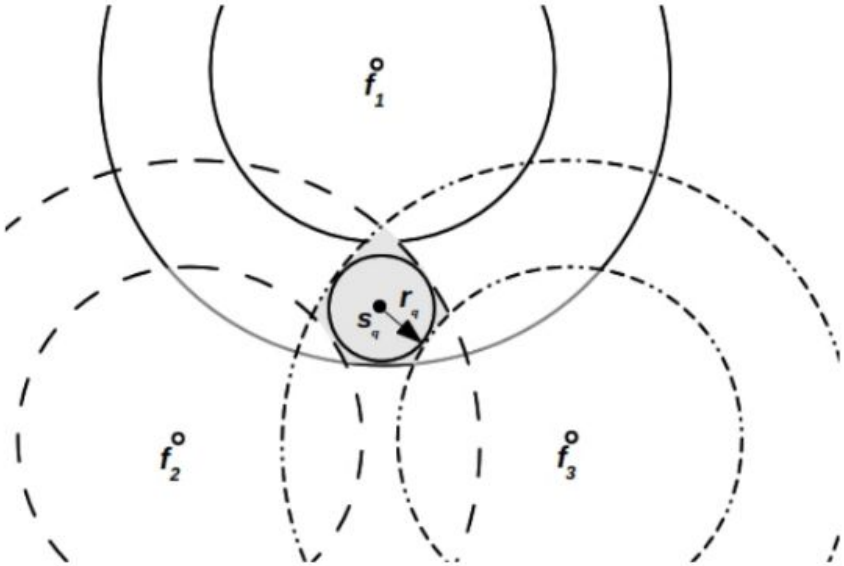
- Dentre os fatores que influenciam na delimitação de uma boa *mbOr* estão a quantidade de representantes globais e o posicionamento deles no espaço métrico.

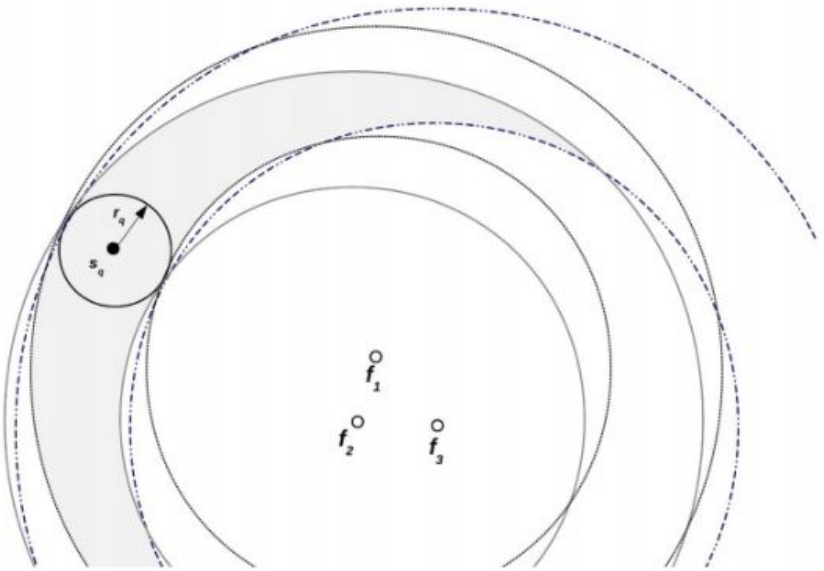
- Um conjunto de elementos representantes é considerado adequado quando ele reduz a *mbOr* o máximo possível, diminuindo a quantidade de distâncias a serem calculadas na etapa de refinamento.

- Na proposta da técnica Omni, foi mostrado que o número de representantes pode ser obtido de acordo com a dimensionalidade intrínseca do conjunto de dados, que consiste da quantidade mínima de atributos necessária para representar os objetos no conjunto.

- Uma maneira de encontrar uma aproximação para a dimensionalidade intrínseca é por meio da correlação da dimensão fractal D_2 (Belussi and Faloutsos, 1995; Traina-Jr et al., 2000).

A Família Omni para Métodos de Acesso Métrico





Obrigado!
Perguntas?

Guilherme Felipe Zabot

Orientador: Profº. Dr. Caetano Traina Jr.