SIMILARITY SEARCH - Poznámky k starým otázkam

1. Metrický priestor, definícia a 3 príklady

Definícia:
metrický priestor M = (D,d)
D - doména dát
d - funkcia vzdialenosti d: D x D -> R(reálne čísla)

Vlastnosti:
Non-negativity
V x,y patrí D, d(x,y) >= 0
Symmetry
V x,y patrí D, d(x,y) = d(y,x)
Identity
V x,y patrí D, x=y <=> d(x,y) = 0
Triangle inequality
V x,y,z patrí d, d(x,z) =< d(x,y) + d(y,z)

Príklady:
Pseudo metric
- neplatí positivness, (môže mať mínusové hodnoty?)
Quasi metric
- neplatí symetria, napr. obsahuje jednosmerné cesty
Super metric
- silnejšia podmienka Triangle inequality

2. vyjmenovat 3 typy dotazu a definovat range query a nearest neighbour query

Range query
- all museums up to 2km from my hotel
Nearest neighbot query k-NN
- five closest museums to my hotel
Reverse nearest neighbot query k-RNN
- all hotels with a specific museum as a nearest cultural heritage
Similarity queries
- pairs of hotels and museums which are 5min walk apart

3. pivot filtering: motivace, priklad, definice, filtering s 1 pivotem a se 2 pivoty.

Object-pivot constraint:

$$\left|d(q,p) - d(p,o)\right| \leq d(q,o) \leq d(q,p) + d(p,o)$$

Pivot filtering
- využíva Triangle inequality pre pruning(prerezávanie)
- poznáme vzdialenosti všetkých objektov k pivotu
- checkujeme len tie objekty, ktoré spĺňajú podmienku

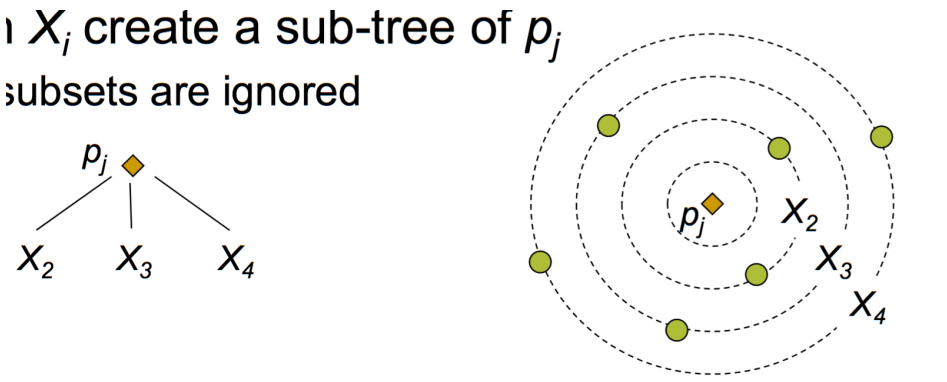
4. precision, recall

Precision
- presnosť je pomer medzi vrátenými kvalifikovanými objektmi a všetkými vrátenými objektmi

Recall
- pomer medzi vrátenými kvalifikovanými objektmi a celkovým počtom kvalifikovaných objektov

5. Burkhard-Keller Tree

- funkcia vzdialenosti musí byť diskrétna
- rekurzívne delí dataset X
- pivot je koreň stromu, subsety podľa vzdialeností



6. m-tree vlastnosti a idea

- dynamická štruktúra, fixná veľkosť uzlov(disk-oriented)
- vytvára sa Bottom-up, zdola nahor
- každý uzol je ohraničený kružnicou, jeho deti sú v rámci kružnice
- iba v listoch sú dátové objekty, tiež ich vzdialenosť k pivotu, ktorý je uložený v rodičovi
- slim tree- varianta so zmenšenými overlapmi

7. Ball partitioning, hyper-plane partitioning.

Ball partitioning
vzdialenosť dm, máme inner set d(q,x)=< dm,
outer set d(q,x) < dm

8. Hyper-plane partitioning

- na základe dvoch pivotov, ku ktorému je x bližšie, tam patrí

9. Levenstein distance

- minumum potrebných úkonov na zmenu stringu a na string b
- úkony insert, delete, replace

10. Minkowskeho vzdialenosti

- tiež nazývané Lp metriky
- definované na n dimenzionálnych vektoroch

$$\sqrt{\sum_{i=1}^{n}(q_i - p_i)^2}.$$ - minkowsky

euklid. - $$\left(\sum_{i=1}^{n}|x_i - y_i|^p\right)^{1/p}$$

11. Jaccardov koeficient

- meranie vzdialenosti medzi setmy A a B

$$d(A,B) = 1 - \frac{|A \cap B|}{|A \cup B|}$$

12. excluded middle partitioning

to isté ako ball partitioning, len vynechá vrstvu okolo obvodu

13. Vantage point tree

Vyberieme pozíciu p v priestore, a rozdelíme na tie, ktoré sú k p bližšie ako treshhold, a tie ktoré nie sú. Opakovane túto procedúru pouýijeme a delíme dáta na menšie a menšie sety. Susedia v strome budú pravdepodobne aj susedia v priestore.
- binárny vyvážený strom

M-tree naco je dobry, ake su tam typy uzlov co je v nich ulozene
leaf node splitting a metody, jakymi se to dela, rozdiel medzi splitovanim v Slim a M-tree

14. D-index

- Distance index kombinuje pivot-filtering a partiotioning(rozdeľovanie)
- viac úrovňová štruktúra založená na hasovaní
- 1.level rozdelí všetky dáta, 2.level rozdelí zvyšnú zónu, posledný level zvyšná zóna ide do zvyškového bucketu celej štruktúry