Znalosti 2008

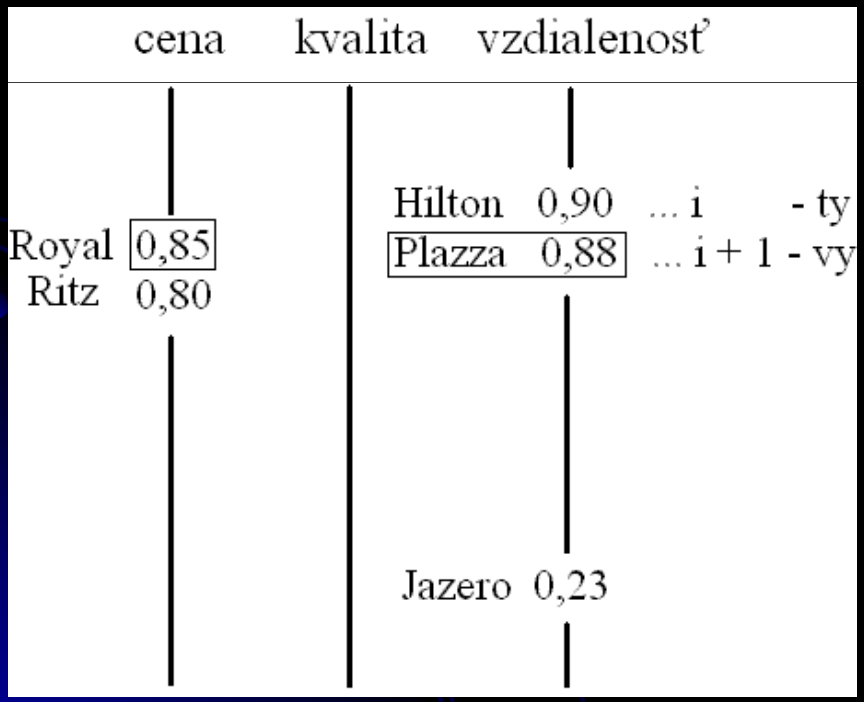
Martin Šumák, Peter Gurský

Preferenčné vyhľadávanie v metrických priestoroch

Ústav informatiky, PF UPJŠ, Košice

pôvod problému: vyhľadávanie najlepších *n* objektov (napr. hotelov)

dáta sú distribuované zoznamy hotelov zotriedené podľa normovaných hodnôt (t.j. z int. <0;1>) hotelov v príslušných atribútoch



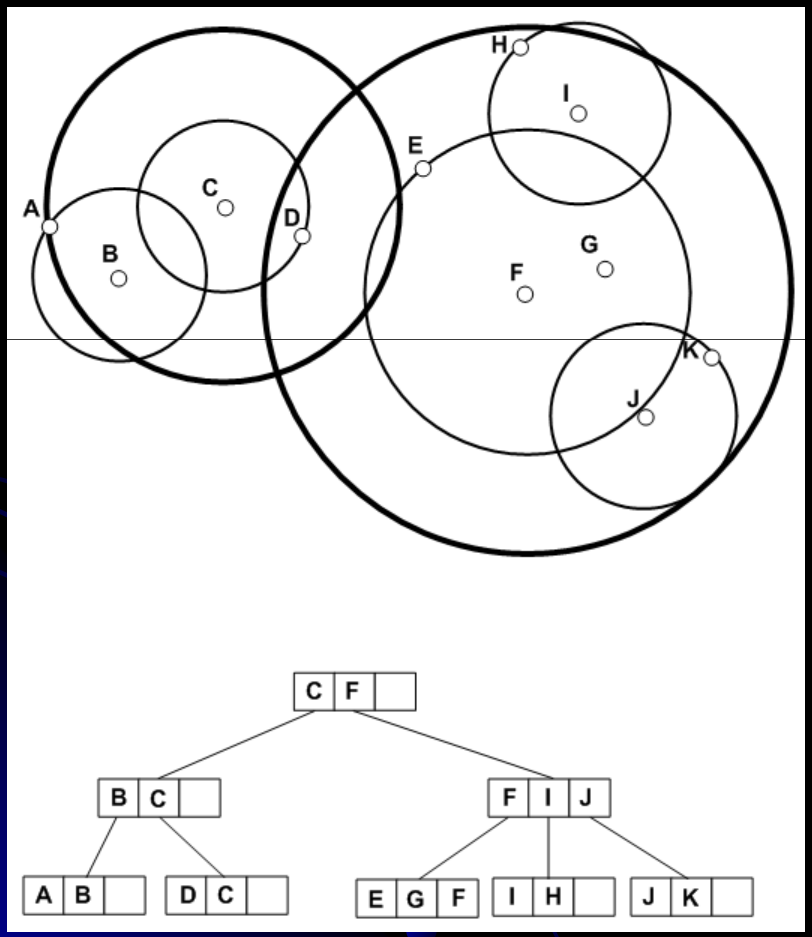
- **cena** – zotriediteľné vopred
- **kvalita** – zotriediteľné vopred
- **vzdialenosť** – !!! zotriediteľné až v čase otázky !!!

predpoklady vs. požiadavky k indexovacej štruktúre

- objekty vieme reprezentovať bodmi v metrickom priestore
- vieme počítať vzdialenosť medzi bodmi
- vzdialenostná funkcia je metrika

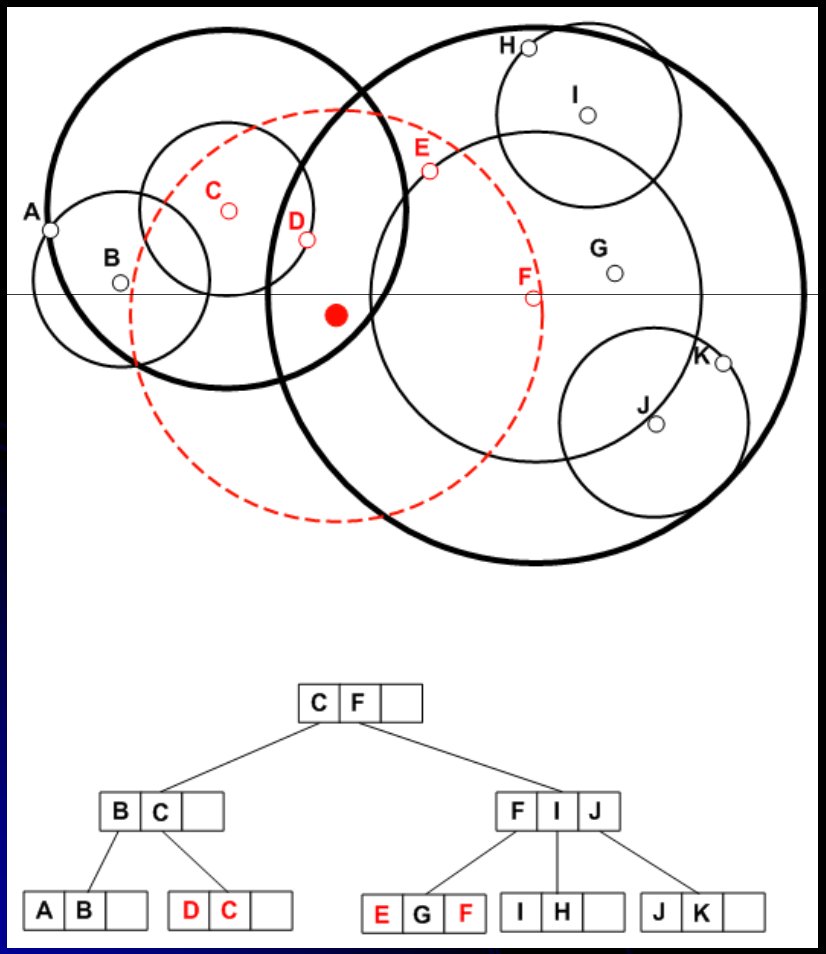
- nezávislosť od použitého metrického priestoru
- nezávislosť od uchovávaných objektov
- používanie sekundárnej pamäte - disk

M-strom

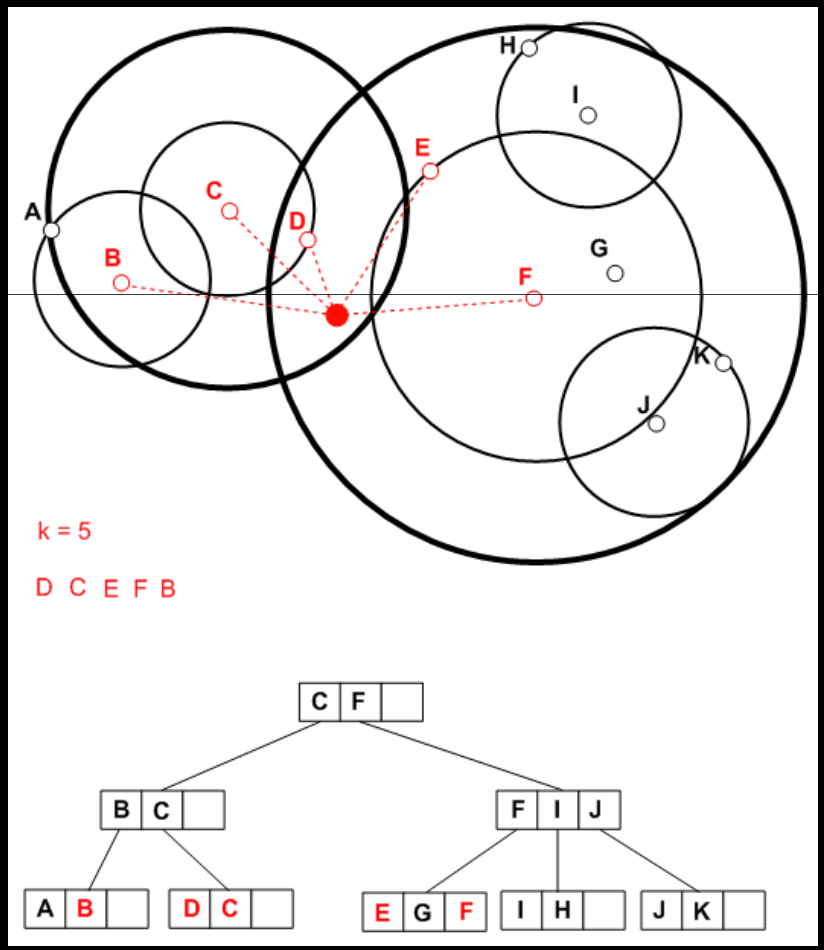


- vyžaduje iba vzdialenostnú funkciu medzi objektami, ktorá je metrikou
- je navrhnutý pre uloženie uzlov na disku
- je vyvážený (podľa vzoru B-stromu)
- poskytuje dynamické vytváranie (podľa vzoru B-stromu)

range query

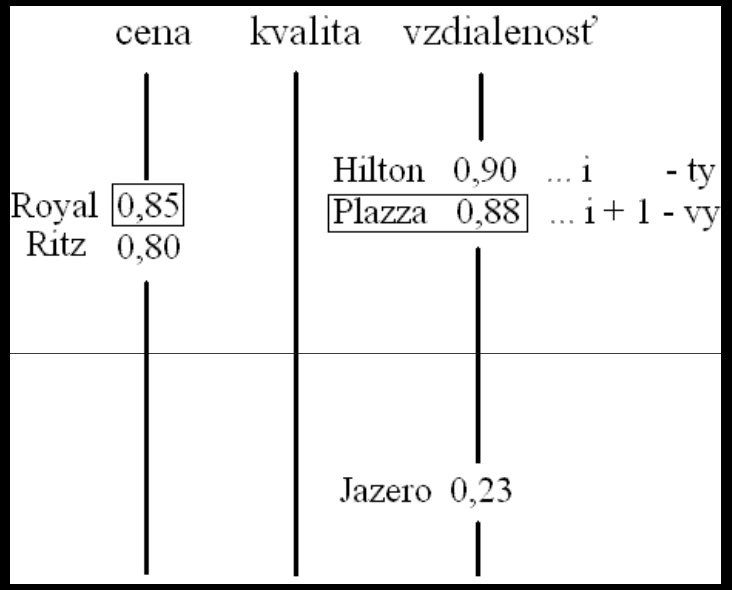


k-nn query



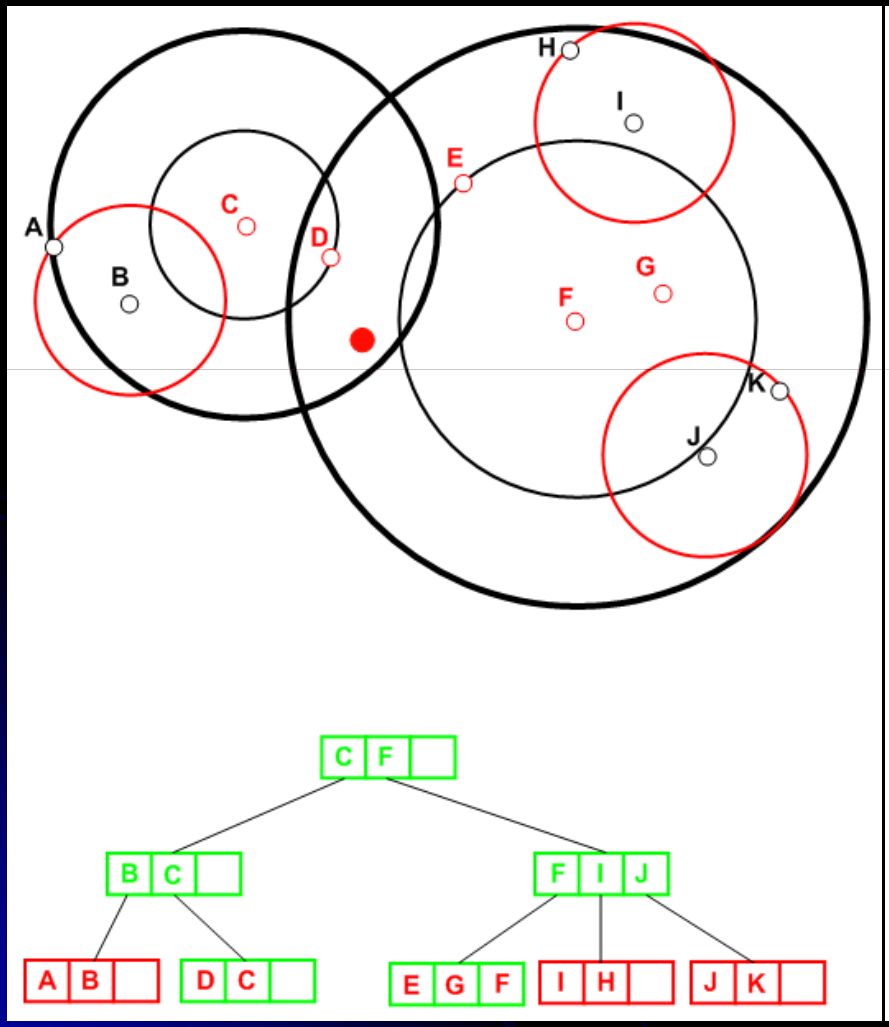
sort query

- objekty vráti v poradí od najbližšieho po najvzdialenejší
- bez obmedzenia *k*, ak treba, vráti všetky objekty v strome



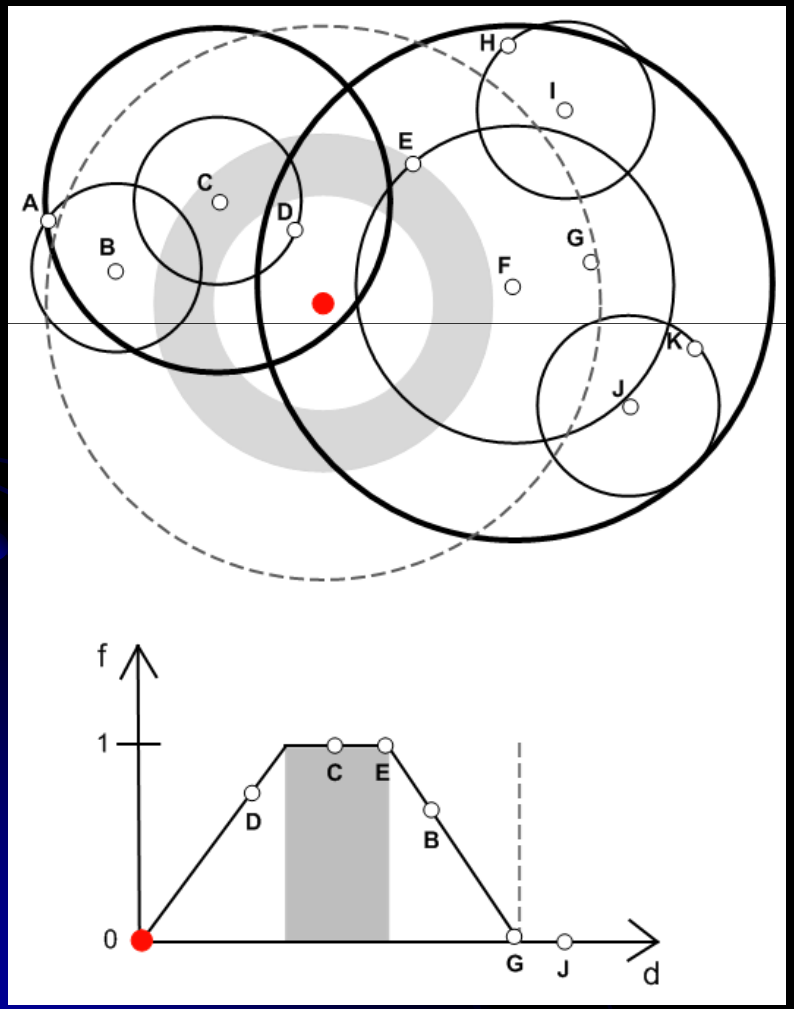
- pre každých *k* najbližších objektov, ktoré vráti, prečíta z disku minimálny nutný počet uzlov (zhodne s *k*-nn query)

činnosť sort query



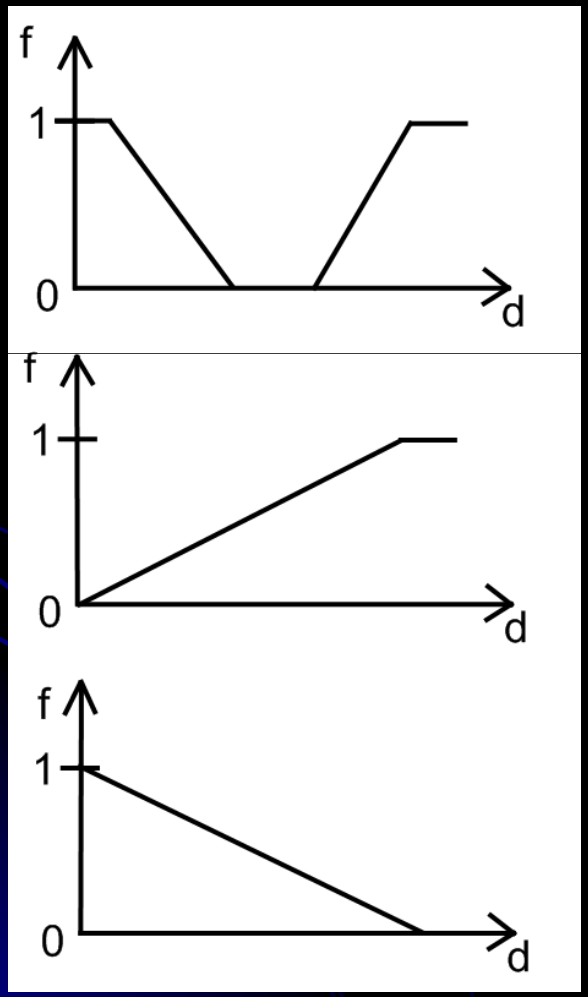
krok	Zoznam čakajúcich **objektov/uzlov** usporiadaný podla minimálnej vzdialenosti	I/Os
0.	C F	0
1.	B C F I J	1
2.	F I J D C A B	2
3.	E G F D C A B I H J K	3
4.	D C A B E F I H J K G	4
5.	D A B C E F I H J K G	5
6.	A B C E F I H J K G	5

fuzzy sort query (pre ohodnotené vzdialenosti)



- nezávislosť od použitej ohodnocovacej funkcie
- funkcia musí byť definovaná na intervale <0, ∞)
- f: <0, ∞) → <0, 1>

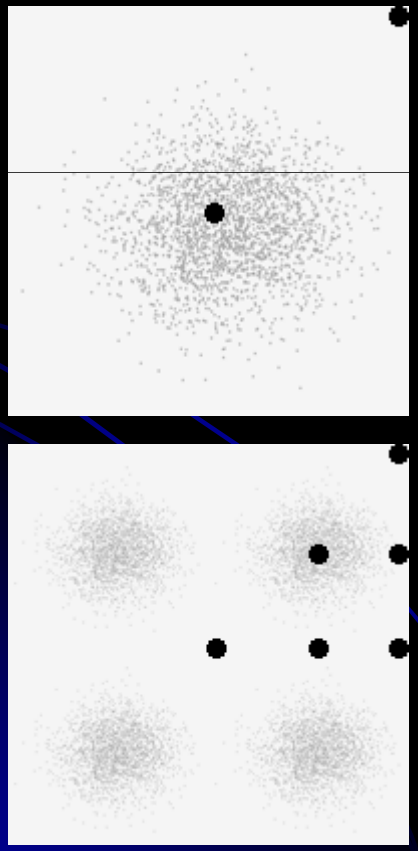
príklady ďalších fuzzy funkcií

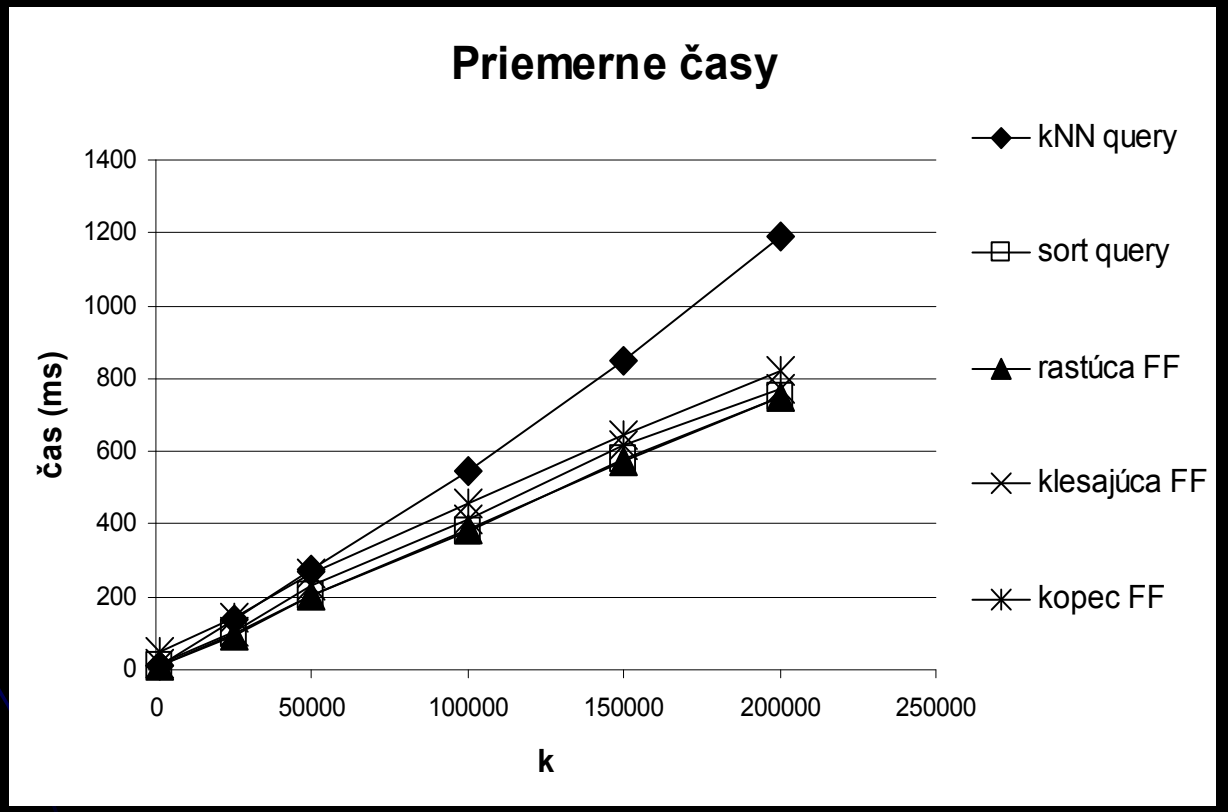


- preferencia blízkych alebo ďalekých objektov
- preferencia ďalekých objektov
- preferencia blízkych objektov

Testy

Ani rôzna distribúcia dát, ani rôzne fuzzy funkcie nespôsobili významné rozdiely vo výkone jednotlivých dopytov!





kNN vs. sort query

- kNN používa okrem pracovného zoznamu aj výsledný zoznam
- pre veľké *k* je udržiavanie obidvoch zoznamov časovo náročnejšie

ďakujem za pozornosť