


Evropský sociální fond
Praha & EU: Investujeme do vaší budoucnosti

Katedra softwarového inženýrství, Fakulta informačních technologií,
České vysoké učení technické v Praze

**VYHLEDÁVÁNÍ NA WEBU A V MULTIMEDIÁLNÍCH DB (BI-VWM)**



# Projekt V - 5

INDEXOVÁNÍ – M-STROM

RNDr. David Hoksza, PhD.

## ZADÁNÍ

Cílem projektu je vytvoření vlastní implementace metrické přístupové metody M-strom.

### VSTUP

Rozsahový nebo kNN dotaz.

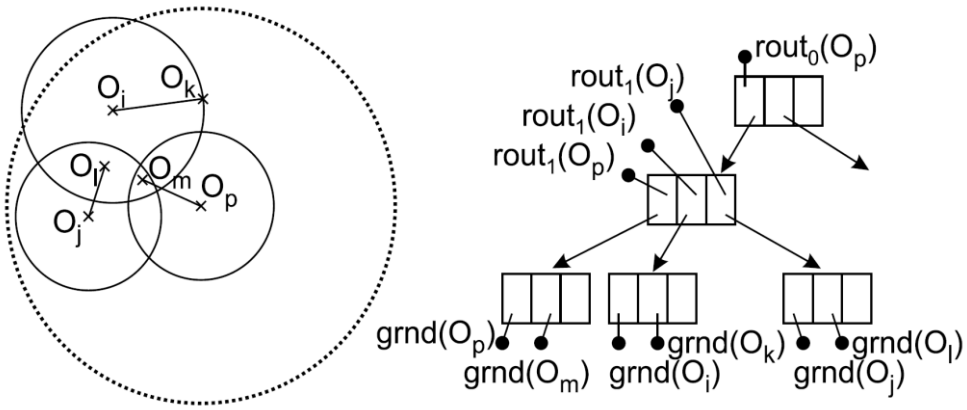
### VÝSTUP

Seznam databázových objektů odpovídající dotazu.

## INFORMACE/POTŘEBNÉ ZNALOSTI

Pro efektivní vyhledávání v prostorech, kde je vzdálenost mezi dvojicí objektů vyjádřitelná funkcí splňující axiomy metriky, se využívají tzv. metrické přístupové metody (MAM). MAM využívají vlastnosti metrické funkce, především trojúhelníkovou nerovnost, k odfiltrování irelevantních objektů bez nutnosti počítat vzdálenost mezi dotazem a irelevantním objektem. Pro odfiltrování je použit spodní odhad vzdálenosti mezi dotazem *D* a objektem *O* na základě znalosti vzdáleností mezi *D* a jiným databázovým objektem *P* (spočítané dříve v průběhu vykonání dotazu) a vzdálenosti a mezi *P* a *O* (spočítané v průběhu dotazu nebo předpočítané). Spodní odhad pak vzniká z uvedených vzdáleností aplikací trojúhelníkové nerovnosti.

Jedna z nejvýznamnějších skupin MAM jsou hierarchické (stromové) MAM. Nejznámější metodou této třídy MAM je M-strom. Podobně jako mnoho struktur v jiných oblastech indexování, i struktura M-stromu je založena na myšlence B+-stromu, tj. je vyvážená, dynamická a stránkovaná (umožňující efektivní perzistenci). Konkrétní index M-stromu představuje hierarchii metrických regionu (každý uzel představuje jeden region), respektive hierarchii shluků objektů v těchto regionech. Listy M-stromu obsahují záznamy *grnd(Oi)* (ground entries) samotných indexovaných objektů *Oi*, zatímco vnitřní uzly obsahují tzv. směrovací záznamy *rout(Oj)* (routing entries). Směrovací záznamy popisují tzv. metrické regiony, které vymezují v metrickém prostoru oblast, v níž se nacházejí objekty uložené v listech příslušného podstromu. Metrický region je popsán hyper-koulí se středem v nějakém objektu a příslušným pokrývajícím poloměrem (covering radius). Příklad hierarchie metrických regionu (pro Euklidovskou vzdálenost a 2D prostor) a příslušného M-stromu je uveden na obrázku.




Indexování objektu v M-stromu je realizováno pouze pomocí dané metriky d. Díky tomu je snadné implementovat dva základní typy dotazu na podobnost. Je to jednak rozsahový dotaz, který slouží k nalezení takových objektu, pro které je vzdálenost od objektu dotazu menší než daná prahová hodnota a dále dotaz na k nejbližších sousedů (k-NN dotaz), kterým získáme prvních k nejméně vzdálených objektů od objektu dotazu. Vyšší efektivita (rychlost) vyhledávání v M-stromu (vůči např. prostému sekvenčnímu průchodu

množiny objektů) spočívá v postupném odfiltrování těch větví M-stromu, které obsahují (vzhledem k dotazu) irelevantní objekty. Korektnost filtrování zaručuje zejména axiom trojúhelníkové nerovnosti metriky.

Je třeba si uvědomit, že zrychlení je závislé i na vzdálenostní funkci, která definuje vzdálenost mezi dvojicí objektů. MAM typicky vycházejí z předpokladu výpočetně drahé vzdálenostní funkce, a tedy jsou cílené na minimalizaci výpočtu vzdálenostních operací. Čím dražší výpočet vzdálenosti je, tím větší zisk oproti sekvenčnímu průchodu MAM vykazují.

Dalším důležitým faktorem je distribuce objektů v prostoru. Jsou-li data špatně klastrovaná, pak není možné vytvořit hierarchii, kde jsou pokrývající regiony dobře separované a tudíž je při dotazování nutné projít velikou část podstromu. Tím se pak výhoda indexu ztrácí a v nejhorším případě může být i horší než sekvenční průchod.

## STAVBA APLIKACE

Aplikace by měla obsahovat:

- Indexaci datové sady pomocí metody M-strom.
- Rozsahové a kNN dotazování pomocí metody M-strom.
- Jednoduché GUI pro možnost testování metody.

## POZNÁMKY K ŘEŠENÍ

Metoda by měla být perzistentní, tj. měla by existovat možnost jejího uložení na disku. Dále by měla metoda umožňovat pracovat i tehdy, když se celý index nevejde do paměti.

## DATA

Data můžou být libovolná. Buď lze použít reálnou databázi nějakých produktů s mnoha dimenzemi, nad kterými se definuje nějaká podobnostní funkce (např. euklidovská vzdálenost), nebo lze vygenerovat náhodnou datovou sadu.

## EXPERIMENTY

V tomto projektu se nabízejí experimenty na testování zrychlení při použití indexu oproti sekvenčnímu průchodu, kdy je objekt porovnáván s každým objektem z DB. Dále lze testovat zrychlení s ohledem na výpočetní náročnost vzdálenostní funkce, vliv distribuce objektů v prostoru, velikost uzlů, atd.

## ZDROJE


- Přednáška *Úvod do podobnostního vyhledávání v multimediálních databázích.*
- Přednáška *Indexování metrické podobnosti pro rychlé vyhledávání v multimediálních databázích.*
- P. Ciaccia, M. Patella, and P. Zezula. M-tree: *An Efficient Access Method for Similarity Search in Metric Spaces.* In Proceedings of the 23rd Athens Intern. Conf. on VLDB, pages 426–435. Morgan Kaufmann, 1997.
- M. Patella. *Similarity Search in Multimedia Databases. PhD thesis.* Dipartmento di Elettronica Informatica e Sistemistica, Bologna, http://www-db.deis.unibo.it/Mtree/index.html, 1999.