

Čtení č. 1 – Fast Approximate Nearest Neighbors with Automatic Algorithm Configuration

student: **Petr Buchal**, xbucha02

Příprava na čtení

Uveďte alespoň 3 dotazy k danému tématu čtení a případným nejasnostem. Pokud je vše jasné, uveďte věci, které sami pokládáte za nejdůležitější pro pochopení problematiky a ostatní by je určitě měli vědět.

- Jakým způsobem funguje Nelder-Mead downhill simplex metoda?
- Co je to GNAT tree?
- Lze nějakým způsobem zredukovat počet dimenzí, tak aby mohla být aproximace nejbližších sousedů využita na tomto redukovaném prostoru a byla rychlejší za cenu snížené přesnosti danou menším počtem dimenzí?

Poznámky ze čtení

Zformulujte své poznatky ze čtení. K čemu jsou tyto metody dobré? Kdy je která metoda vhodná? Co mi v článku připadá silné a co slabé? Co stále ještě nechápu? Alespoň 5 bodů.

- Kd-stromy jsou neefektivní pro používání na více než 3 dimenzích.
- V praxi se kd-strom konstruuje tak, že se dělí dimenze, která má největší rozptyl. V článku vybrali 5 dimenzí s největším rozptylem a z těchto pěti dimenzí vybrali jednu náhodně.
- Parametry metody v článku jsou počet kd-stromů a počet bodů. Počtem bodů se ovlivňuje přesnost. Počet kd-stromů ovlivňuje rychlost a paměťovou náročnost. Počet stromů se hledá pomocí optimalizace zadaných parametrů – paměťová a časová náročnost.
- Autoři článku používají hierarchický k-means, ten na výsledek k-means opět aplikuje k-means a ukládá podstromy. Zároveň u k-means tree používají prioritní frontu, kterou plní nejbližším stromem, který nenavštívili.
- Hierarchický k-means strom je totéž co vocabulary tree.
- Zrychlení je větší s klesající požadovanou přesností.

Kontrolní otázky

Stručně odpovězte na kontrolní otázky uvedené na Wiki stránce.

1. Co je to problém hledání intervalů (Range Searching) a jaké má praktické využití?
 - Mějme množinu bodů v prostoru, problém hledání intervalů spočívá ve zpracování a uložení bodů do takových struktur, které v okamžiku obdržení intervalu by měli vrátit rychle všechny body, které se v něm nachází.
2. Jaké výhody má datová struktura Skip Tree a jak se do ní vkládá nová hodnota/uzel?
 - Nepracuje se s celým stromem, ale je to lokální operace -> možnost paralelních operací. Hlavní výhodou je jednoduché mazání a vkládání uzlů. Ukládání uzlu: zamknutí, vložení, odemknutí.
3. Jak probíhá tvorba kd-tree?
 - Data se rozdělí podle mediánu rovnoběžného s osou x, všechny data nalevo od rozdělovací hranice jsou potomci levého podstromu kd-stromu a potomci napravo od rozdělovací hranice jsou potomci pravého stromu. Takhle se prostor dělí rekurzivně dále s tím, že se střídají osy.
4. Jak lze definovat tzv. Approximate Nearest Neighbors Search?
 - Jedná se o aproximaci nejbližších sousedů bodu v n-rozměrném prostoru.
5. Jaké dvě základní datové struktury pro ANN search používá knihovna FLANN?
 - Randomized kd-tree, hierarchický k-means tree.

Doporučení pro budoucí čtení

Co vám vyhovovalo a co nevyhovovalo na čtení? Přineslo vám čtení něco pozitivního? Které prvky by měly být zachovány, zesíleny, potlačeny, eliminovány? Alespoň jeden bod.

- Domluvit se na tématu článku se studenty.