Oponentský posudek diplomové práce Matúše Roštára

Zpracování proudů dat

Obsah práce

Zpracování proudů dat je charakteristické velkým objemem dat, která mohou být čtena pouze jednou. Uložení dat pro jejich pozdější využití nebo opakované čtení je limitováno dostupnou pamětí. Předkládaná diplomová práce se zaměřuje na vyhodnocení dlouho běžících dotazů nad proudy dat.

Text je rozdělen do tří základních částí — teoretický úvod, vlastní přínos a testy. Teoretický úvod pokrývají první dvě kapitoly, kde jsou vysvětleny základní pojmy, popsán obecný systém řízení proudů dat a základní metody pro efektivní zpracování dotazů nad proudy dat s využitím omezené paměti. Vlastní přínos práce pokrývají kapitoly 3 a 4. Kapitola 3 obsahuje návrh algoritmu shlukování dat s využitím histogramu pro proudy dat. Kapitola 4 se věnuje návrhu a implementaci prototypu systému pro vyhodnocování dotazů a dále implementaci navrženého algoritmu. Experimentálním testům jsou věnovány kapitoly 5, 6 a 7, v nichž autor srovnává navržený algoritmus s dalšími dvěma zvolenými algoritmy. Poslední kapitola shrnuje celou práci. K diplomové práci je přiloženo CD s implementovaným prototypem systému pro zpracování proudů dat.

Hlavní klady práce

Z poměrně obecného zadání si autor vybral a specifikoval úzkou část, které se v práci věnoval. Jedná se o zpracování dlouhoběžících dotazů. Z práce samotné jsou dobře napsány zejména kapitoly 1 a 2, které dávají čtenáři přehled o obecné architektuře systémů pro zpracování proudů dat i podrobnější popis metod používaných pro výběr dat reprezentujících celý proud v omezené paměti.

Autor navrhl algoritmus, který pro vyhodnocení dotazu nad proudem dat využívá dostupnou paměť k ukládání statistických informací o proudu tak, aby dotaz mohl být vyhodnocen pouze na základě uložených informací co nejpřesněji vzhledem k obsahu proudu dat. Hlavní předností navrženého algoritmu je jeho přesnost vyhodnocení dotazu nad velkým objemem dat v omezené paměti, která byla otestována a popsána v kapitole 6. Z popisu lze očekávat, že algoritmus dosahuje dobré přesnosti i při malé paměti dostupné pro statistická data.

Podle dostupné literatury nebyl podobný algoritmus ještě nikdy otestován ani implementován.

Věcné nedostatky

Teoretický návrh algoritmu (kapitola 3) a dále i popis implementace (kapitola 4) jsou zmatené. Z celé práce není jasně vidět, jak vypadá celkový algoritmus, který kombinuje posuvné okno, shlukování a histogram. Autor nejprve v podkapitole 2.6 podrobně vysvětlí algoritmy z literatury [16], ale později v kapitole 3 na straně 23 píše: „V tejto práci je však na zhlukovanie použitý algoritmus [19] s využitím [20] ako podprocedúrou." Ale algoritmus z literatury [19] není nikde jinde v práci popsán. V podkapitole 4.3 na straně 37 se dočteme, že nakonec byly implementovány algoritmy z [16] i [20], ale v testech se bude používat jen ten z [16].

Nikde v kapitolách 5 a 6, v nichž jsou popsány experimentální testy, nejsou uvedeny parametry algoritmu, s nimiž byly testy provedeny. Konkrétně se jedná o velikost posuvného okna (parametr N, strana 29), počet „přihrádek" histogramu (parametr m, strana 29), parametr t (strana 30), parametr k (strana 31).

Autor v práci srovnává tři algoritmy s konstatováním, že všechny měly k dispozici stejný prostor pro data. Protože ale nikde v práci nejsou uvedeny parametry algoritmů použitých v testech, nelze toto tvrzení ověřit. Díky tomu ani není jasné, kolik paměti (vyjádřeno např. v počtu hodnot v posuvném okně) měly algoritmy k dispozici, a tedy jestli nejsou výborné výsledky navrženého algoritmu ovlivněny velkým množstvím uložených informací.

Test přesnosti algoritmu, který je „navržen" v podkapitole 6.1, není dle mého názoru vhodný. V testu se totiž srovnávají algoritmy pomocí průměrné odchylky (strana 28, Definice 3) mezi

průměrem z celého proudu dat a průměrem spočítaným testovaným algoritmem po přečtení počáteční úseku proudu. Tímto se ovšem spíše testuje, jak velkou část proudu dat potřebuje algoritmus pro extrapolaci průměru. Test přesnosti by měl porovnávat algoritmy pomocí průměrné odchylky mezi průměrem z počátečního úseku proudu vyhodnoceného přesně (na základě všech hodnot) a vyhodnoceného testovaným algoritmem. Tím by se totiž otestovala přesnost aproximace průměru z velkého množství dat s využitím omezené paměti. Navržený algoritmus by díky tomu předčil ostatní testované algoritmy ještě výrazněji.

Strana 58 se zdá být nedokončená.

V kapitole 3.2 autor popisuje, jak se slučují „přihrádky" histogramu, ale nepíše „kdy" se slučují. To je sice vidět v kapitole 3.5, ale tam zase není uvedeno „proč" to tak je.

Autor v závěru práce algoritmus hodnotí jako pomalý, ale v možnostech jeho zlepšení navrhuje pouze jednu metodu zlepšení.

Přiložené zdrojové kódy obsahují minimum komentářů. Zejména se jedná o špatně zdokumentované atributy tříd a parametry funkcí.

Jako závažný nedostatek považuju i typografickou kvalitu předložené diplomové práce, která je na velmi nízké úrovni. Hlavní prohřešky jsou jednopísmenné spojky a předložky na koncích řádků, velmi nízká úroveň sazby matematiky — formátování vzorců, nekonzistence v používání matematického fontu (např. strany 26, 28, 38), šedě podbarvený obsah (strany 3, 4), nekonzistentní formátování zdrojových kódů a algoritmů (strany 20, 32, 40, 41), volné poloviny stránek 47 a 50.

Méně závažné nedostatky a připomínky

Čtení textu komplikuje použití stejného označení pomocí písmen v různých významech v rámci jedné kapitoly (strany 28, 29), či dokonce ve stejném odstavci (strana 31 dole).

Přesnost a rychlost navrženého algoritmu mohla být srovnána pro různá nastavení jeho parametrů.

Autor v podkapitole 7.4 navrhuje nasazení implementovaného algoritmu např. při zpracování proudu dat z bezpečnostních videokamer, které mají snímkovací frekvenci 1 snímek za sekundu. Rád bych jen upozornil, že algoritmus podle testů dokáže zpracovat přibližně 3 čísla za sekundu, nikoliv 3 obrázky za sekundu.

V podkapitole 5.1 jsou data testovací sady 1 charakterizována jako „Prvá testovacia sada obsahuje prvky pseudonáhodne rozmiestnené v spektre 0 až 100 000, kde pseudonáhodnosť je zabezpečená funkciou `random()` z jazyka Java." Bylo by lepší uvést také jméno balíku a třídy, protože Java nemá přímo žádnou funkci `random()`.

Závěr

I přes výše uvedené nedostatky a připomínky předložená diplomová práce splnila zadání a doporučuji ji k obhajobě.

V Praze dne 16. ledna 2008