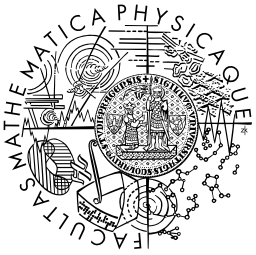

MATEMATICKO-FYZIKÁLNÍ FAKULTA
Univerzita Karlova

BAKALÁŘSKÁ PRÁCE

Martin Gora

Vylepšení agregace dotazovacího enginu pro grafové databáze

Katedra softwarového inženýrství

Vedoucí bakalářské práce: Mgr. Tomáš Faltín
Studijní program: Informatika
Studijní obor: Softwarové a datové inženýrství

Praha 2021

Prohlašuji, že jsem tuto bakalářskou práci vypracoval(a) samostatně a výhradně s použitím citovaných pramenů, literatury a dalších odborných zdrojů. Tato práce nebyla využita k získání jiného nebo stejného titulu.

Beru na vědomí, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorského zákona v platném znění, zejména skutečnost, že Univerzita Karlova má právo na uzavření licenční smlouvy o užití této práce jako školního díla podle §60 odst. 1 autorského zákona.

V ..... dne ....

Podpis autora

Chtěl bych poděkovat mému vedoucímu Mgr. Tomáši Faltínovi za jeho pomoc, ochotu a nadšení při zpracovávání daného tématu. Déle bych chtěl poděkovat rodině, která mi poskytla zázemí pro práci a plnou podporu.

Název práce: Vylepšení agregace dotazovacího enginu pro grafové databáze

Autor: Martin Gora

Katedra: Katedra softwarového inženýrství

Vedoucí bakalářské práce: Mgr. Tomáš Faltín, Katedra softwarového inženýrství

Abstrakt: Abstrakt.

Klíčová slova: grafové databáze agregace dat proudové systémy

Title: Improvement of data aggregation in query engine for graph databases

Author: Martin Gora

Department: Department of Software Engineering

Supervisor: Mgr. Tomáš Faltín, Department of Software Engineering

Abstract: Abstract.

Keywords: graph databases data aggregation streaming systems

Obsah

Úvod

Tady ma byt text.

1. prvni

aaa Anděl (2007, Věta 4.22) aa

2. druha

3. treti

4. Experiment

Aby bylo možné porovnat stávající řešení s nově navrženým řešením na poli rychlosti zpracovávání dotazů, podrobili jsme jsme zmíněná řešení experimentu. Vykonaný experiment proběhne na reálných grafech různé velikosti s uměle vygenerovanými vlastnostmi naležící vrcholům. Nad danými grafy provedeme vybrané množství dotazů, které nám umožní sledovat a porovnat chování řešení v různých situacích.

4.1 Příprava dat

Pro náš experiment jsme použili tři orientované grafy z databáze SNAP[1].

	#Vrcholů	**#Hran**
Amazon0601	403394	3387388
WebBerkStan	685230	7600595
As-Skitter	1696415	11095298

Tabulka 4.1: Vybrané grafy použité k experimentu

- **Amazon0601:** Jedná se o graf vytvořený procházením webových stránek Amazonu na základě featury „Customers Who Bought This Item Also Bought“ ze dne 1.6.2003. V grafu existuje hrana z i do j, pokud je produkt i často zakoupen s produktem j.
- **WebBerkStan:** Graf popisuje odkazy webových stránek domén `https://www.stanford.edu/` a `https://www.berkeley.edu/`. Vrcholem je webová stránka a hrana představuje hypertextový odkaz mezi stránkami.
- **As-Skitter:** Topologický graf internetu vytvořený programem `traceroute` z roku 2005. Ačkoliv je uvedeno, že daný graf je neorientovaný, vnitřní hlavička souborů uvádí opak, proto jsme se daný graf rozhodli přesto využít.

Samotné grafy obsahují pouze seznam hran. Následuje ukázka souboru grafu Amazon0601:

```
# Directed graph (each unordered pair of nodes is saved once):
    Amazon0601.txt:
# Amazon product co-purchaisng network from June 01 2003
# Nodes: 403394 Edges: 3387388
# FromNodeId        ToNodeId
0        1
0        2
0        3
0        4
0        5
0        6
0        7
```

[1] Leskovec a Krevl (2014)

4.1.1 Transformace grafových dat

Abychom mohli dané grafy použít, museli jsme je přetransformovat do platného formátu vstupních souborů dotazovacího enginu. Výstupem transformace budou soubory popisující schéma vrcholů/hran NodeTypes.txt/EdgeTypes.txt a samotné datové soubory vrcholů/hran Nodes.txt/Edges.txt. V našem případě graf bude obsahovat pouze jeden typ hrany a jeden typ vrcholu. Dané omezení ovlivňuje pouze vyhledávání vzoru, které není přínosné pro náš experiment, protože snižuje počet nalezených výsledků.

Cílem transformace je získat formát datových souborů Nodes.txt/Edges.txt odpovídající schématu:

```
Soubor EdgeTypes.txt:
[
{
"Kind": "BasicEdge"
}
]

Soubor NodeTypes.txt:
[
{
"Kind": "BasicNode"
}
]
```

Program provádějící transformaci je obsahem přílohy zdrojových kódů A.1 v souboru GrapDataBuilder.cs, který vygeneruje datové soubory Nodes.txt a Edges.txt dle výše vypsaného schématu. Pro připomenutí zmíníme, že prvni sloupeček v datových souborech Edges.txt a Nodes.txt odpovídá unikátnímu `ID` v rámci celého grafu. Výstupní soubor Edges.txt bude obsahovat hrany setřízené v rostoucím pořadí dle položky `FromNodeId` s přidělenými `IDs` od hodnoty `ID` posledního vrcholu v souboru Nodes.txt. Samotný soubor Nodes.txt obsahuje setřizené vrcholy podle `ID` v rostoucím pořadí. Je nutné zmínit, že setřízení dat podle `ID` není nežádoucí, jelikož nezaručuje nic o seskupení vrcholů v daném grafu.

Následuje ukázka výstupních souborů transformace pro graf Amazon0601:

```
Soubor Edges.txt:
403395 BasicEdge 0 1
403396 BasicEdge 0 2
403397 BasicEdge 0 3
403398 BasicEdge 0 4
...

Soubor Nodes.txt:
0 BasicNode
1 BasicNode
2 BasicNode
3 BasicNode
...
```

4.1.2 Generování Properties vrcholů

Posledním krokem přípravy dat pro experiment je vygenerovat Properties vrcholů. Jsme si vědomi, že nejideálnější způsob testování je graf s reálnými daty, nicméně dané omezení jsme se rozhodli aplikovat kvůli problematickému hledání vhodných dat s triviální transformací do vhodného vstupního formátu. Proto pro každý vrchol náhodně vygenerujeme hodnoty tří Properties.

Property	Type	Popis
PropOne	integer	`Int32` s rozsahem $[0, 100000]$
PropTwo	integer	`Int32` s rozsahem [`Int32.MinValue`, `Int32.MaxValue`]
PropThree	string	délka $[2, 8]$ ASCII znaků s rozsahem $[33, 126]$

Tabulka 4.2: Generované Properties vrcholů

- **PropTwo** hodnoty jsou rovněž generováný střídavě kladně a záporně, aby nastal rovnoměrný počet záporných a kladných hodnot.
- **PropThree** hodnoty jsou pouze ASCII znaky z rozsahu $[33, 126]$. Dané omezení výplývá z vlastností dotazovacího enginu, aby bylo možné bez obtíží načíst datový soubor.

Na základě tabulky generovaných Properties 4.2 následuje ukázka upraveného souboru schématu pro vrcholy:

```
Soubor NodeTypes.txt:
[
{
"Kind": "BasicNode",
"PropOne": "integer",
"PropTwo": "integer",
"PropThree": "string"
}
]
```

Výsledné hodnoty Properties do souborů Edges.txt/Nodes.txt jsou vygenerovány pomocí programu, který používá generátor náhodných čísel. Program je obsažen v příloze zdrojových kódů A.1 v souboru PropertyGenerator.cs. Pro každý graf bylo použité jiné `Seed` pro inicializaci náhodného generátoru v daném programu. Samotná `Seeds` byla vygenerována rovněž náhodně.

	Seed
Amazon0601	429185
WebBerkStan	20022
As-Skitter	82

Tabulka 4.3: Inicializační hodnoty náhodného generátoru pro PropertyGenerator.cs

Program generuje hodnoty definované ve statické položce `propGenerators` a zachovává jejich pořadí ve výsledném datovém souboru. Aby nedocházelo k omylům při opakování experimentů, uvádíme útržek kódu použité inicializace položky dle tabulky generovaných vlastností 4.2 pro všechny tři grafy:

```
static PropGenerator[] propGenerators = new PropGenerator[]
{
    new Int32Generator(0, 100_000, false),
    new Int32Generator(true),
    new StringASCIIGenerator(2, 8, 33, 126)
};
```

Timto jsme dokončili poslední nutný krok k vygenerování platných vstupních dat pro dotazovací engine. Použité grafy k transformaci a výsledné datové soubory jsou obsahem přílohy grafů pro experiment A.3

4.2 Výběr testovaných dotazů

V této sekci provedeme výběr a popis dotazů, které budou objektem testování.

4.3 Metodika

C

4.3.1 HW

4.4 Výsledky a diskuze

Závěr

Tady ma byt text

Seznam použité literatury

Anděl, J. (2007). *Základy matematické statistiky.* Druhé opravené vydání. Matfyzpress, Praha. ISBN 80-7378-001-1.

Leskovec, J. a Krevl, A. (2014). SNAP Datasets: Stanford large network dataset collection. `http://snap.stanford.edu/data`.

Seznam obrázků

Seznam tabulek

Seznam použitých zkratek

A. Přílohy

A.1 Zdrojové kódy

Přílohou této bakalářské práce jsou zdrojové kódy dotazovacího enginu, benchmarku a použité knihovny HPCsharp. Vše zmíněné je přiloženo v rámci jednoho projektu Visual Studia, kromě souborů Gitu. Dále, mimo projekt jsou přiloženy zdrojové kódy programů na generování vstupních grafů pro experiment. Jedná se o soubory GraphDataBuilder.cs a PropertyGenerator.cs.

A.2 Online Git repozitář

V době vydání tohoto textu probíhal vývoj dotazovacího enginu na GitHubu.

`https://github.com/goramartin/QueryEngine`

A.3 Použité grafy při experimentu

Grafy použité při experimentu jsou vloženy do odpovídajících složek dle názvu grafu. Složky obsahují originálni grafy před transformací a datové soubory po transformaci.

A.4 druha priloha

Priloha po prvni strance priloh