Indexování struktur v grafovém DB stroji Neo4j

Autor: Ing. Martin Troup | Vedoucí práce Ing. Michal Valenta, Ph.D | Škola: ČVUT - Fakulta informačních technologií

Úvod

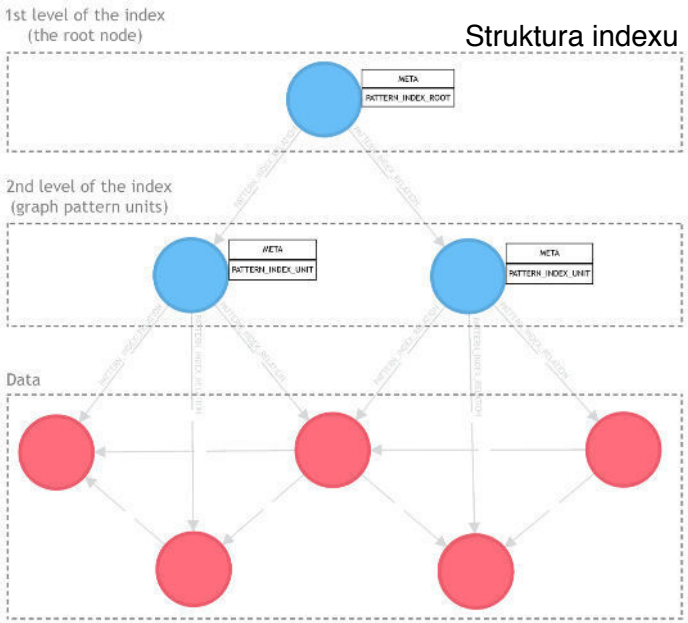
Grafové databáze patří do kategorie NoSQL databází.Jde o velmi novou a moderní technologii, která se na trhu vyskytuje od roku 2003. Grafové databáze ukládají data nativně do grafového modelu a tedy efektivně využívají vlastností grafu. Narozdíl od typických relačních databází je zde kladen velký důraz na vztahy mezi jednotlivými datovými objekty. Právě díky optimalizované práci se vztahy mezi objekty lze efektivně provádět různé operace, které jsou v relačních databázích velmi náročné, především pak kvůli drahé operaci JOIN.

Motivace

Z praxe vychází mnoho případů užití, které vyžadují vyhledávání konkrétních grafových vzorů v grafové databázi. Jde například o detekci vzorů v sociálních sítích pro doporučovací systémy, detekci transakčních podvodů, které jsou popsatelné specifickými vzory podezřelého použití kreditních karet v transakčních sítích, či například detekce bodů selhání v telekomunikačních sítích. Všechny tyto a další mnohé případy použití grafových databází sdílejí stejný problém, a to hledání podgrafů v grafu. Tato operace je velmi náročná, protože je nutné projít celou databázi, respektive graf a z každého uzlu pak dohledávat potenciální grafové vzory pomocí prohledání do šířky či hloubky. Tato komplexní operace je pro grafové databáze klíčová a proto vzniká prostor pro její zefektivnění. Již dlouhou dobu se v relačních databází používají indexy k rychlému dohledávání záznamů a proto je logickým krokem k zefektivění vyhledávání vzorů právě indexace.

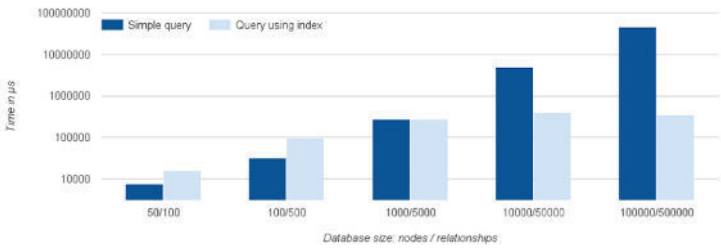
Implementace

V práci byla navrhnuta kompletní struktura indexu mapující grafové vzory a zároveň byl navrhnut způsob jak index udržet aktuální při běžné práci s databází, tedy při aplikování DML operací. Součástí práce je spolu s návrhem také implementace plně funkčního prototypu systému indexování grafových vzorů pro momentálně nějvětší implementaci grafových databází, Neo4j. Tento prototyp lze ovládat pomocí REST API a lze ho do Neo4j databáze integrovat jako samostatný modul. Imlementace byla provedena v jazyku Java, který je spolu se Scalou hlavním stavebním kamenem databáze Neo4j. Byl aktivně využit framework GRAPHAWARE, který je po samotném rozhraní Neo4j dalším možným ovládacím prvkem databáze. Práce byla po celou dobu konzultována s autorem frameworku, Michalem Bachmanem, MSc. a částečně s týmem Neo4j.



Měření

Bylo provedeno testování a porovnávání efektivity dotazování grafových vzorů v databázi, respektive vyhledávání podgrafů v grafu. Pro měření byl vytvořen benchmark, ve kterém se ukázalo, že dotazy s podporou indexu uvedeném v této práci jsou rychlejší o řád až dva řády podle typu dotazovaného grafového vzoru a typu a množství dat v databázi. Jde tedy o poměrně velký krok k optimalizaci operace dotazování na vyhledání grafových vzorů v grafové databázi, která je, jak je výše zmíněno, klíčová pro mnoho případů užití této databáze.



Závěr

V práci se podařilo navrhnout a implementovat první prototyp pro indexování grafových struktur v grafové databázi. Využití indexu vykazuje poměrně velké zefektivnění o proti stávajícím možnostem dotazování.

Zdroje

[1] Robinson, Webber, Eifrem. Graph Databases. ISBN 1449356265 , 2013

[2] Bachman. GraphAware: Towards Online Analytical Processing in Graph Databases. Imperial College London, 2013

[3] Tvrdík. Parallel algorithms and computing. CVUT, 2005