Úvod k bakalářské práci

Vladimír Míč

11. prosince 2012

S rozvojem informačních technologií dochází k obrovskému nárůstu velikosti ukládaných a zpracovávaných dat. Společně s tím dochází i k rozvoji sítí, a tím se zlepšuje datová dostupnost a možnosti sdílení. Hlavně v síťové oblasti se dnes běžně setkáváme se servery, které obsahují řádově desítky terabytů dat. Tento objem zaplní zejména multimediální soubory, konkrétně obrázky a videa, která se dnes po síti velmi rychle šíří. Pro zachování praktické použitelnosti, tedy dostatečně rychlé reakce na běžné příkazy, je nutné vylepšovat hardware a zároveň zavádět do praxe nové metody pro práci s daty [1].

Tato bakalářská práce vznikla pro potřeby vyhledávacího systému. Vyhledávání v rozsáhlých množinách dat už není možné realizovat lineárním průchodem jejich celého objemu, jako je tomu u menších datových úložišť, ani indexováním veškerého obsahu [2]. Data proto musí být předzpracována. Vhodnou volbou je použití některého z algoritmů pro shlukování. Jedná se o algoritmy, které roztřídí objekty do množin (dále třídy nebo shluky) a to tak, že objekty ve stejné třídě si jsou co nejvíce podobné a objekty v rozdílných třídách jsou co nejvíce odlišné [3]. Pro potřeby vyhledávání je dále vybrán pro každou třídu reprezentant. Ten se v rámci dané podobnosti nachází blízko středu, což znamená, že vystihuje danou třídu nejlépe. Reprezentantem může být i fiktivní objekt, který leží přesně uprostřed, a vystihuje tak danou třídu lépe, než kterýkoli její reálný objekt. Vyhledávání se realizuje pouze nad jednotlivými reprezentanty, čímž dochází k urychlení procesu. Algoritmy provádějící hierarchické shlukování zavádí více úrovní shluků. Počet úrovní může a nemusí být omezen.

Cílem bakalářské práce bylo implementovat v jazyce Java vybrané algoritmy hierarchického shlukování množiny obecných bodů a porovnat jejich výsledky, zejména rychlost a kvalitu zpracování. Požadované bylo využití knihovny Messif[1] , což je knihovna jazyka Java vyvíjená na Fakultě informatiky Masarykovy univerzity v Brně (dále FI MU) pro práci nad metrickými prostory. Program měl být schopen zpracovávat různé typy dat. Podmínkou pro ně byla pouze jejich reprezentace pomocí třídy v jazyce Java, a možnost jejich porovnávání s jiným objektem stejného typu pomocí metody getDistance(). Výsledný program měl být schopen rozdělit zpracovávání dat paralelně na více počítačů tak, aby bylo možné konečný výsledek jednoduše získat z dílčích výsledků. Každý počítač měl také zpracovávat úlohu paralelně v rámci vícevláknové aplikace. Tím mělo být dosaženo co nejvyšší efektivity a rychlého zpracování při použití v praxi. Při psaní programu bylo využito třídy AbstractObject a dalších tříd knihovny Messif, čímž byly splněny první dva požadavky. Program využívá paralelismu na úrovni vláken a knihovnu Hadoop[2] společnosti Apache pro rozdělení výpočtů

[1]Stránky knihovny Messif: http://lsd.fi.muni.cz/trac/messif
[2]Stránky knihovny Hadoop: http://hadoop.apache.org

mezi více počítačů. Tím byly splněny zbývající požadavky.

Program byl začleněn do vyhledávacího systému Mufin[3], rovněž vyvíjeného na FI MU, který slouží jako univerzální vyhledávací systém. V době vzniku bakalářské práce byl zaměřen zejména na obrázky, ale počítalo se s jeho použitím v rámci podobnostního vyhledávání v biologii, geografii, při vyhledávání plagiátů a v dalších oblastech. Matematický popis konkrétních dat, například obrázků, a tvorba metody getDistance() nebyly úkolem bakalářské práce.

Práce obsahuje i kapitolu věnující se testování konkrétních algoritmů hierarchického shlukování. V této kapitole jsou vidět rozdíly mezi časovou náročností algoritmů a kvalitou shluků při aplikaci na konkrétní soubor jednoho milionu objektů, jimiž byly textové popisy obrázků z galerie Flickr.com. Tyto popisy byly vytvořeny v rámci projektu Mufin. V praxi to může být i tisíckrát tolik. Po dohodě s vedoucím práce bylo usouzeno, že pro potřeby systému Mufin bude nejvhodnější MacNaughton–Smithova metoda hierarchického shlukování.

Bakalářská práce obsahuje přehled základních metod hierarchického shlukování a jejich popis. Dále jsou uvedeny a rozebrány konkrétní algoritmy hierarchického shlukování včetně jejich složitostí, návrh programu, jeho implementace a testování pro použití pěti různých shlukovacích algoritmů. Diskutován je problém zastavení algoritmů.

První kapitola je věnována metodám hierarchického shlukování. Druhá kapitola se zabývá konkrétními algoritmy, třetí pak návrhem výsledného programu, čtvrtá jeho implementaci a pátá testováním.

Rejstřík

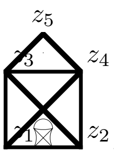
Reference

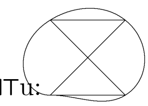
[1] Výzkum a vývoj v oblasti flexibilních infrastruktur.

[2] Mark Baggesen. A review of windows search 4.0 from microsoft, 2011.

[3] Jiří Kučera. Shluková analýza, 2008.

Grafika

Znak v METAPOSTu: 

, znak v METAFONTu: 

[3]Stránky projektu Mufin: http://mufin.fi.muni.cz/tiki-index.php