Distribuované zhlukovanie textových dokumentov v prostredí Gridu

Martin Sarnovský, Peter Butka, Vladimír Safko

FEI-KKUI/FEI-CIT TU Košice

Obsah

- Úvod
- Ciele a motivácia
- Algoritmus GHSOM
- Návrh distribuovanej verzie
- Použité nástroje
- Implementácia
- Experimenty a výsledky
- Závery a zhodnotenie

Využitie Gridu pri objavovaní znalostí v dátach a textoch

- V súčasnosti existuje niekoľko prostredí pre podporu distribuovaného dolovania dát na Gride
  - TeraGrid
  - DataMiningGrid
  - KnowledgeGrid
  - Discovery-Net
  - GridMiner
  - UK National Centre for TextMining
- Grid v sebe integruje distribuované a paralelné počítanie
- Vedie aj ku vzniku nových algoritmov a techník

Motivácia

- Zhlukovane textových dokumentov zhlukovacie algoritmy a možnosti ich distribúcie
- Algoritmus GHSOM (Growing Hierarchical Self Organizing Map)
- Návrh distribúcie algoritmu GHSOM a jeho implementácia v integro-vanom prostredí pre distribuované dolovanie textových dokumentov na Gride - GridMiner

GHSOM

- Samoorganizujúce sa mapy (Self-Organizing Maps, SOM) - metóda sekvenčného nehierarchického zhlukovania
  - Princíp nekontrolovaného konkurenčného učenia
  - Mapuje vysoko-rozmerný príznakový priestor do dvojrozmerného priestoru - mapa
  - Nevýhody algoritmu SOM
    - štruktúra mapy musí byť definovaná apriori
    - výsledné mapy sú príliš veľké a je problém s nimi pracovať
- GHSOM (Growing Hierarchical Self Organizing Map), mapa rastie:

  hierarchicky – podľa distribúcie dát, čo umožňuje hierarchickú dekompozí-ciu a navigáciu v podmapách
  horizontálne – veľkosť mapy sa mení tak, aby sa prispôsobila požiadavkám vstupného priestoru

Distribuovaný Algoritmus

- Distribuovaný algoritmus GHSOM je implementovaný v jazyku Java ako služba systému GridMiner na gridovej vrstve tohto systému s využitím knižnice JBowl
- Po vytvorení mapy GSOM 1. úrovne vzniká niekoľko samostatných zhlukovacích procesov
  - budovanie hierarchických podstromov GHSOM pozostávajúcich z hierarchicky usporiadaných máp GSOM
  - paralelné vykonávanie týchto zhlukovacích procesov na uzloch

Distribuovaný Algoritmus

- Prístup Master-Worker (Hlavný uzol-Pracovné uzly)
- Na hlavnom uzle sa vyráta celková odchýlka vstupných dát, vytvorí sa mapa 1. úrovne, označia sa neuróny, ktoré spĺňajú podmienku expanzie do podmapy
- Zo vstupných vektorov sa vyberú vektory prislúchajúce neurónom, ktoré sa budú expandovať a rozpošlú na jednotlivé uzly
- Na uzloch sa vektory stanú vstupom pre algoritmus GHSOM, ktorý vytvorí hierarchický podstrom
- Na hlavnom uzle sa prijaté čiastkové modely GHSOM spoja do jedného výsledného hierarchického modelu GHSOM
- Podmienky ukončenia rastu GHSOM
  - hĺbka hierarchie = max. definovaná hĺbka
  - neexistuje centroid na ďalšiu expanziu

Použité nástroje

- GridMiner
  - Framework pre distribuované dolovanie dát a OLAP na výpočtovom Gride
  - Vybudovaný na Globus toolkit 3.0 (v súčasnosti sa prechádza na verziu 4.0) – gridové služby
  - 3 vrstvy:
    - gridová vrstva
    - webová vrstva
    - grafické užívateľské rozhranie
- JBowl
  - Java Bag of Words Library
  - Open-source knižnica napísaná v jazyku Java
  - Poskytuje API pre predspracovanie a analýzu textových dokumentov

Experimenty

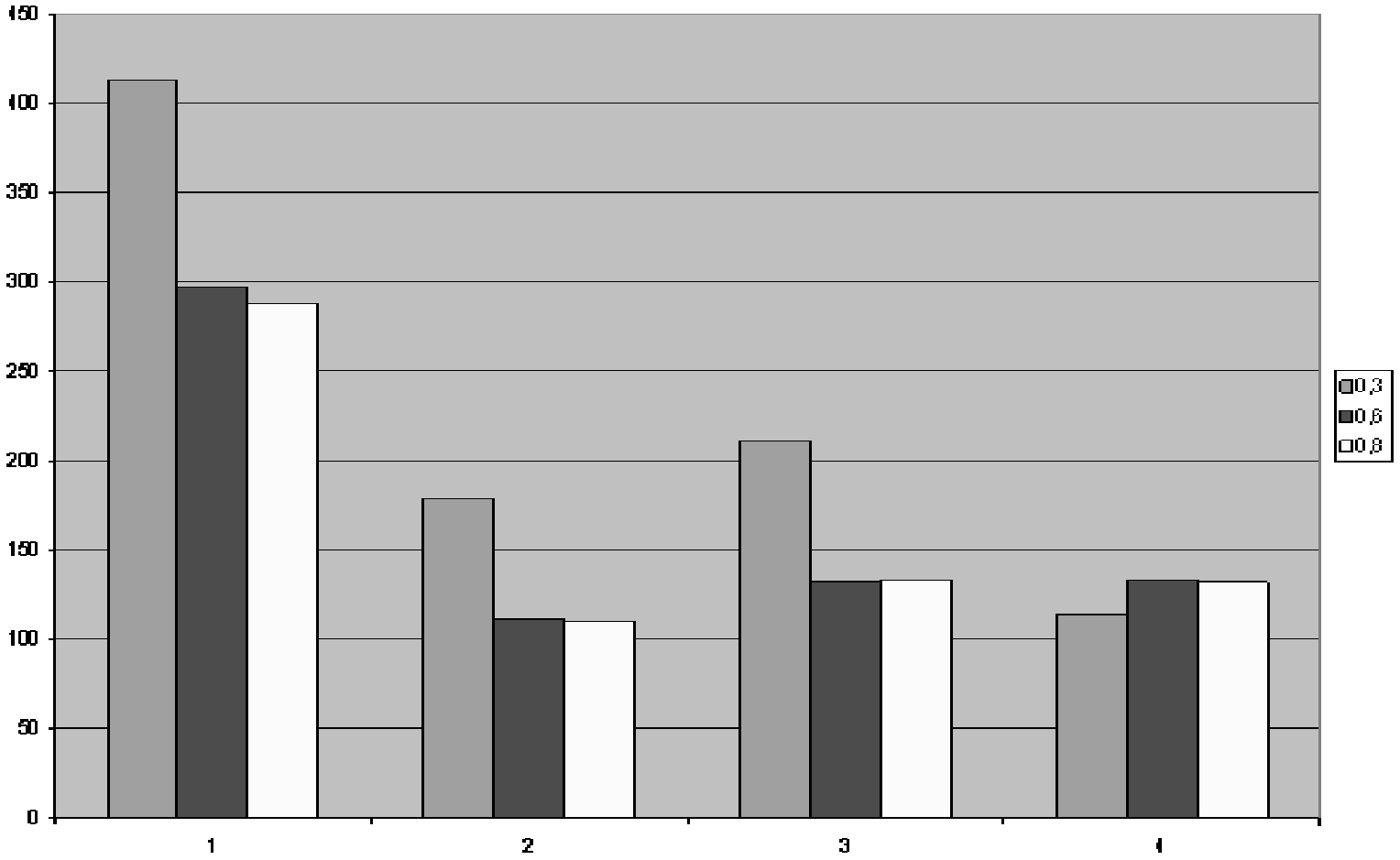
- Cieľom experimentov - porovnať časovú náročnosť sekvenčného algoritmu GHSOM s jeho distribuovanou verziou
- V experimentoch sa menil počet uzlov a parameter *tau1*, ktorý riadi výslednú kvalitu zhlukov, čím ovplyvňuje veľkosť vytváraných máp GSOM a celkovú hĺbku hierarchie modelu GHSOM
- Použité dátové množiny
  - Times DataSet (420 dokumentov)
  - Reuters DataSet (7769 dokumentov)

Testovacie Prostredie

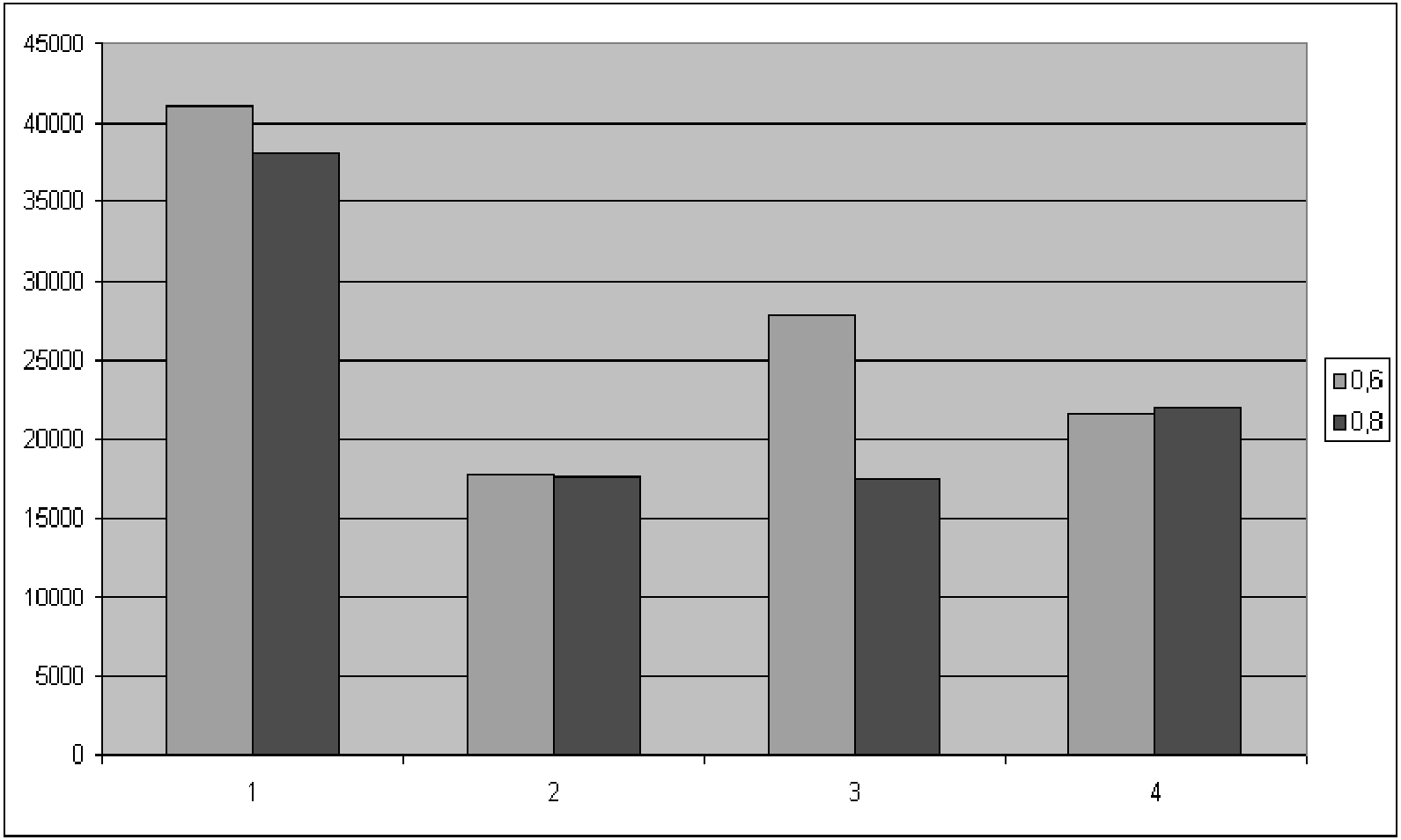
- Distribuovaná verzia bola testovaná v testovacom prostredí pozostávajúceho zo servera (4 x UltraSPARC-III 750MHz, 8GB RAM), ktorý predstavoval hlavný uzol
- Pracovné stanice SUN, ktorých konfigurácia je v nasledujúcej tabuľke
- Uzly Gridu boli prepojené sieťou 100Mbit/s

Node	Configuration
3	Sun Fire X 4600 M2 (Solaris x86), 32GB RAM
4	Sun-Blade-1500 (UltraSPARC-IIIi 1062MHz), 1.5GB RAM
2	Sun-Fire-V445 (4x Ultra SPARC- IIIi 1592MHz), 16384MB RAM
1	Sun-Fire-V240 (2x Ultra SPARC- IIIi 1503MHz), 12288MB RAM

Experiments on Times Dataset



Experiments on Reuters Dataset



Závery a Zhodnotenia

- Výsledky ukazujú celkové urýchlenie algoritmu GHSOM pri jeho distribuovanej verzii
- Nerovnomerné rozdelenie dát na jednotlivé uzly a tiež veľkosť chyby, ktorú daný uzol znižoval,
  - Zobrať do úvahy aj výkon jednotlivých uzlov Gridu: centroidy s najväčším počtom dokumentov priradiť najvýkonnejším uzlom Gridu
- Rôzny výpočtový výkon jednotlivých uzlov testovacieho prostredia
- Ďalšou možnosťou ako rovnomernejšie a efektívnejšie využívať uzly Gridu, by mohla byť taká distribúcia algoritmu GHSOM, pri ktorej by sa na uzloch Gridu vytvárali iba mapy GSOM a nie celé podstromy GHSOM

Vďaka za pozornosť

???