Automatische Kompressorwahl für SCIL

Armin Schaare

Betreuer: Dr. Julian Kunkel, Anastasiia Novikova

Arbeitsbereich Wissenschaftliches Rechnen
Fachbereich Informatik
Fakultät für Mathematik, Informatik und Naturwissenschaften
Universität Hamburg

2016-07-13

Zeitplan, letzte 2 Wochen

- Festlegung, kompressionsrelevante Größen
- Extraktion relevanter Datencharakteristiken
- Benchmarken von Kompressionsverfahren und Konfigurationen

Festlegung, kompressionsrelevater Größen

- Unterteilung in:
  - Benutzerspezifizierte Kompressionsinfos
  - Platformabhängige Benchmark-Resultate
  - Charackteristiken der zu komprimierenden Daten

Benutzerspezifizierte Kompressionsinfos

- Absolute Fehlertoleranz
- Relative Felhertoleranz
- Kompressionsdurchsatz
- Dekompressionsdurchsatz

Platformabhängige Benchmark-Resultate

- Kompressionsdurchsatz
- Dekompressionsdurchsatz
- Kompressionsrate

Charakteristiken der zu komprimierenden Daten

- Durchschnitt
- Standardabweichung
- Minimum
- Maximum
- Datengröße

Charakteristiken der zu komprimierenden Daten

- 10-dim. Suchraum, deswegen:
- Sinvoll den Prozess aufzuteilen
- Aufteilung nach Durchsatz und Rate
- Vorfiltern nach Erwarteten K-Spezifikationen

1. Abschätzung von K-Durchsätzen

- Tatsächliche K-Raten und -Durchsätze zu Anfang nicht bekannt
- Deswegen: Curve-Fit für jede K-Methode
  - Abbildung: abs./rel. und alle Charakteristiken $\rightarrow$ K-Durchsatz, D-Durchsatz
  - 7-dim. Parameter-Suchraum
  - Kompliziertes Problem an und für sich
- Ergebnis: Funktion zum Abschätzen von K-Durchsätzen

2. Abschätzung von K-Raten

- Mit Hilfe der Abschätzung: Vorfiltern von KMs
- Restkandidaten: nach bestmöglicher K-Rate bewerten
- Dazu: Neuer Curve-Fit für jede K-Methode
  - Abbildung: abs./rel. und Charakteriskiken $\rightarrow$ K-Rate
  - wieder 7-dim Parameter-Suchraum
- Ergebnis: Funktion zum Abschätzen von K-Raten

Extraktion relevanter Datencharakteristiken

- Einfach: Berechne Boxplot für zu komprimerende Daten
  - Durchschnitt
  - Standardabweichung
  - Minimum
  - Maximum
  - Datengröße
- Problem: Bei großen Daten zu langsam
- Lösung: Stichprobenartig Werte berechnen

Benchmarken von Kompressionsverfahren und Konfigurationen

- Niederlegen der Ergebnisse im csv-Format
- 7-dim. Suchraum $\rightarrow 10^7$ Gitterpunkte bei nur 10 verschiedenen Werten pro Größe
- Sinnvoll: Jeden Gitterpunkt oft Berechnen, um statistische Fehler zu ermitteln
- Problem: 100 Berechnungen für 100 GP $\rightarrow 10^{28}$ Einträge.
- Was nun?

Erwartete Probleme

- Suchraum zu komplex $\rightarrow$ Zu viele Trainingsdaten nötig
- Optimale Funktion zu komplex $\rightarrow$ Abschätzungen ungenau
- Stichprobenartige Extraktion von Daten ungenau/langsam

Zeitplan, nächste 2 Wochen

- Suchen: Lösung/Tradeoff für Benchmark Explosionsproblem
- Einarbeiten in R
- Erste Testevaluationen