

**SVEUČILIŠTE U ZAGREBU**  
**PRIRODOSLOVNO–MATEMATIČKI FAKULTET**  
**MATEMATIČKI ODSJEK**

Anto Čabraja

**PARALELNI ALGORITMI ZA  
PROBLEM GRUPIRANJA PODATAKA**

Diplomski rad

Voditelj rada:  
prof. dr. sc. Goranka Nogo

Zagreb, srpanj 2014.

Ovaj diplomski rad obranjen je dana \_\_\_\_\_ pred ispitnim povjerenstvom  
u sastavu:

1. \_\_\_\_\_, predsjednik
2. \_\_\_\_\_, član
3. \_\_\_\_\_, član

Povjerenstvo je rad ocijenilo ocjenom \_\_\_\_\_.

Potpisi članova povjerenstva:

1. \_\_\_\_\_
2. \_\_\_\_\_
3. \_\_\_\_\_

# Sadržaj

<b>Sadržaj</b>	<b>iii</b>
0.1 Problem grupiranja podataka . . . . .	1
0.2 Primjena . . . . .	1
0.3 Pregled rada . . . . .	1
<b>1 Modeliranje problema grupiranja</b>	<b>3</b>
1.1 Osnovni pojmovi . . . . .	3
1.2 Matematičko modeliranje problema . . . . .	4
1.3 Metode razvoja algoritama za grupiranje . . . . .	5
1.4 Upravljanje podacima . . . . .	5
<b>2 Metaheuristike</b>	<b>7</b>
2.1 Prirodom inspirirani algoritmi . . . . .	7
2.2 Reprezentacija podataka . . . . .	7
2.3 Analiza rezultata . . . . .	7
<b>3 Poznati algoritmi i analiza</b>	<b>9</b>
3.1 Alg 1 . . . . .	9
3.2 Alg 2 . . . . .	9
3.3 Alg 3 . . . . .	9
<b>4 Tehnike za paralelizaciju algoritama</b>	<b>11</b>
4.1 Osnovni pojmovi MPI tehnologije . . . . .	11
4.2 Topologija . . . . .	11
4.3 Prednosti paralelizacije i cijena komunikacije . . . . .	11
<b>5 Konstrukcija paralelnih algoritama za grupiranje</b>	<b>13</b>
5.1 Algoritam 1 heurisika . . . . .	13
5.2 Algoritam 2 iterativno . . . . .	13
5.3 Algoritam 3 hibrid . . . . .	13

<b>6 Ostale moderne metode</b>	<b>15</b>
6.1 Programiranje na grafičkim karticama . . . . .	15
6.2 MapReduce metoda . . . . .	15
<b>Bibliografija</b>	<b>17</b>

# **Uvod**

## **0.1 Problem grupiranja podataka**

## **0.2 Primjena**

## **0.3 Pregled rada**



# Poglavlje 1

## Modeliranje problema grupiranja

### 1.1 Osnovni pojmovi

Kako bi u daljnjem razmatranju bilo jednostavnije objašnjavati strukture i same implementacije algoritama potrebno je problem grupiranja reprezentirati osnovnim pojmovima. U nastavku ćemo formalno definirati sve komponente od kojih se problem grupiranja sastoji.

**Definicija 1.1.1.** *Uzorak je apstraktna struktura podataka koja reprezentira stvarne podatke s kojima raspolaže algoritam za grupiranje.*

**Definicija 1.1.2.** *Svojstvo je vrijednost ili struktura koja predstavlja jednu značajku danog podatka unutar uzorka.*

**Definicija 1.1.3.** *Udaljenost između uzoraka definiramo kao funkciju  $f : D \rightarrow \mathbb{R}$ , gdje je  $D$  skup svojstava danih uzoraka*

**Definicija 1.1.4.** *Za uzorke kažemo da su **blizu** jedan drugome ako je njihova udaljenost manja od unaprijed zadane veličine*

**Definicija 1.1.5.** *Klaster je skup uzoraka koji su u prostoru podataka blizu. Ako su uzorci identični onda je njihova udaljenost uvijek 0*

**Definicija 1.1.6.** *Jedinstveno grupiranje je postupak grupiranja kada svaki uzorak pripada jednom i samo jednom klasteru.*

**Definicija 1.1.7.** *Nejasno ili nejedinstveno grupiranje je postupak grupiranja gdje jedan uzorak može biti u više klastera.*

**Napomena 1.1.8.** *U radu ćemo promatrati **jedinstveno grupiranje** tako da će sve daljnje definicije i modeliranja pretpostavljati da želimo dobiti disjunktne klastere. Jedinstveno grupiranje (eng: hard clustering) je ujedno i teži problem.*

## 1.2 Matematičko modeliranje problema

Definicija grupiranja podataka nije jedinstvena. U literaturi se na različite načine pokušava opisati ovaj postupak. Neki od pokušaja opisne definicije su:

1. *Grupiranje podataka je postupak otkrivanja homogenih<sup>1</sup> grupa uzoraka unutar skupa svih danih uzoraka.*
2. *Grupiranje podataka je postupak određivanja koji su uzorci slični te ih svrstati u isti klaster.*

Za modelirati problem neće nam biti dovoljne opisne definicije. U ovom slučaju opisne definicije mogu poslužiti samo kao intuicija o čemu se zapravo radi kada govorimo o grupiranju. U nastavku ćemo pomoću definiranih pojmova u poglavlju 1.1 matematički opisati problem grupiranja podataka.

Neka je  $\mathbf{U} = \{U_1, U_2, \dots, U_n\}$  skup od  $n$  uzoraka, te neka je  $U_i = (s_1, s_2, \dots, s_d)$  reprezentiran  $d$ -dimenzionalnim vektorom gdje  $s_i$  predstavlja jedno svojstvo. Ovako definiran  $\mathbf{U}$  moguće je reprezentirati kao matricu  $\mathbf{S}_{d \times n}$ . Svaki stupac te matrice predstavlja jedan uzorak iz danog skupa  $\mathbf{U}$ .

$$\mathbf{S}_{d \times n} = \begin{pmatrix} s_{1,1} & s_{1,2} & \cdots & \cdots & s_{1,n} \\ s_{2,1} & s_{2,2} & \cdots & \cdots & s_{2,n} \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ s_{d,1} & s_{d,2} & \cdots & \cdots & s_{d,n} \end{pmatrix} \quad (1.1)$$

Iz definicije 1.1.6 te iz navedenog formalnog zapisa dajemo formalnu definiciju problema grupiranja.

**Definicija 1.2.1.** *Skup od  $k$  klastera  $\mathbf{C} = \{C_1, C_2, \dots, C_k\}$  je skup sa sljedećim svojstvima:*

- $C_i \neq \Phi$
- $C_i \cap C_j = \Phi, \forall i, j \text{ } i \neq j$
- $\bigcup_{i=1}^k C_i = \mathbf{U}$

**Napomena 1.2.2.** *U terminima matrice  $\mathbf{Z}$  to znači da se svaki  $C_i$  zapravo sastoji od stupaca matrice  $\mathbf{Z}$ .*

**Definicija 1.2.3.** *Problem grupiranja u skup od  $k$  klastera  $\mathbf{C}$  je ekvivalentan problemu da  $\forall c, c' \in C_i$  udaljenost od  $c$  do  $c'$  je manja od udaljenosti  $c$  do bilo kojeg drugog  $c'' \in C_j$   $j \neq i$*

---

<sup>1</sup>podaci koji se ne mogu smisleno separirati



Zapravo problem grupiranja je pronalazak najpogodnije particije za  $\mathbf{C}$  u skupu svih mogućih particija. Prema napomeni 1.2.2 to znači da se zapravo radi o problemu raspodjele  $n$  stupaca matrice  $\mathbf{Z}$  u  $k$  skupova

### **1.3 Metode razvoja algoritama za grupiranje**

### **1.4 Upravljanje podacima**



## **Poglavlje 2**

# **Meta-heuristički pristup problemu**

**2.1 Prirodom inspirirani algoritmi**

**2.2 Reprezentacija podataka**

**2.3 Analiza rezultata**



## **Poglavlje 3**

### **Poznati algoritmi i analiza**

**3.1 Alg 1**

**3.2 Alg 2**

**3.3 Alg 3**



## **Poglavlje 4**

# **Tehnike za paralelizaciju algoritama**

**4.1 Osnovni pojmovi MPI tehnologije**

**4.2 Topologije**

**4.3 Prednosti paralelizacije i cijena komunikacije**





## **Poglavlje 5**

# **Konstrukcija paralelnih algoritama za grupiranje**

### **5.1 Algoritam 1 heurisika**

**Opis**

**Analiza**

### **5.2 Algoritam 2 iterativno**

**Opis**

**Analiza**

### **5.3 Algoritam 3 hibrid**

**Opis**

**Analiza**



## **Poglavlje 6**

### **Ostale moderne metode**

#### **6.1 Programiranje na grafičkim karticama**

#### **6.2 MapReduce metoda**



# Bibliografija

- [1] I. Autor, *Naslov Knjige*, Samizdat, 2052.
- [2] D. E. Dutkay, D. Han, Q. Sun i E. Weber, *Hearing the Hausdorff dimension*, (2009), <http://arxiv.org/abs/0910.5433>.
- [3] S. Kurepa, *Convex functions*, Glasnik Mat.-Fiz. Astr. Ser. II **11** (1956), br. 2, 89–93.
- [4] ———, *Funkcionalna analiza*, Školska Knjiga, 1981.



# Sažetak

Ukratko ...





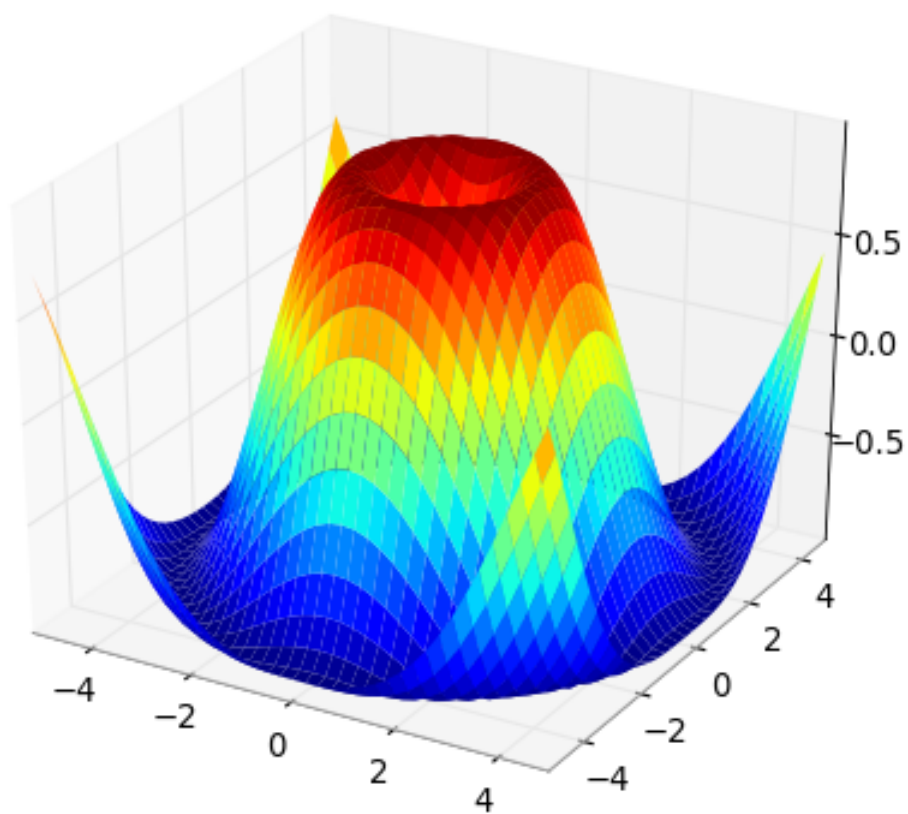
# Summary

In this ...



# Životopis

Na slici 1.1.7 se nalazi 3D graf neke funkcije.



Slika 6.1: Druga slika

kao i jedna vrlo komplicirana formula koja slijedi iz (??)

$$\sum_{i=1}^\infty A_{x_1} \times A_{\alpha_2} \oslash \iint_\Omega x^2 \ddagger \limsup_{n \in \mathbb{N}} \frac{\alpha + \theta + \gamma}{n^\omega} \text{ je u stvari } \bigcup_{r \in \mathbb{Q}} \overline{\Xi_i \ominus_{\substack{j \in \mathbb{C} \\ j \ni i \mathbb{Q}}} \Upsilon^{kj} \Psi \hbar}_{*|\{ \alpha \}}.$$