Research in Databases
Data, Information and Beyond
Mirella M. Moro, UFMG
MIRELLA M. MORO

CONTACT INFORMATION

Mirella M. Moro
Departamento de Ciência da Computação
Universidade Federal de Minas Gerais - UFMG
Av. Antônio Carlos, 6627
Pampulha, Belo Horizonte - MG, Brazil
CEP 31270-010
Office: Building ICEX, room 4012
Phone numbers:
+55 (31) 3409-5849 (Office)
+55 (31) 3409-5858 (Fax)
Email: mirella [AT] dcc [DOT] ufmg [DOT] br
Notes

• TMI per minute
• Slides are online
• Numbered slides: questions at the end

Last text only slide – on purpose
- DBMS (database management systems)
- Modeling, design, queries
- Traditional (boring!) applications

**Past**

**NoSQL**: text, documents, graphs & etc

**Connection**: Web, streams, mobile, parallel, workflow, ...

**Context**: cloud, hardware, privacy, provenance...

**Now**

- Collect, extract, index, process, deliver
- E.g., recommendation systems and classifiers

**IR**

**KDD**

- Knowledge Discovery
- Data Mining
- E.g., machine learning e big data analytics

**Databases**
• DBMS (database management systems)
• Modeling, design, queries
• Traditional (boring!) applications

Past

• NoSQL: text, documents, graphs & etc
• Connection: Web, streams, mobile, parallel, workflow, …
• Context: cloud, hardware, privacy, provenance…

Now

• Collect, extract, index, process, deliver
• E.g., recommendation systems and classifiers

IR

• Knowledge discovery
• Data mining
• E.g., machine learning & big data analytics

KD
I’m not the only one

26. Enhanced Data Models for Advanced Applications
27. Introduction to Information Retrieval and Web Search
28. Data Mining Concepts

26. Data Mining
27. Information Retrieval and XML Data
28. Spatial Data Management
29. Further Reading Mobile, Main Memory DB, Info Vis, …

6th ed 2011
3rd ed 2002
Doing research

Prepare to know by heart the next slide
When you are

• Thinking
• Planning
• Writing
• Presenting
• Discussing
• Meeting
• Arguing
- General context
- Specific context [state of the art]

- What does *not* work?
- What can be better?

- Contribution
- How to solve the problem

- Is good? Is better?
Contributions everywhere!!

- **Context**
  - Known X > new context

- **Problem**
  - Known X > novel problem

- **Solution**
  - Known X > new solution
  - Known solution > improved

- **Evaluation**
  - Different evaluations > known X
Contributions everywhere!!

**Context**
- Relational data > Streams
- Relational data > Mobile

**Problem**
- Relational data > semi-structured
- Relational data > big data analysis

**Solution**
- Most common

**Evaluation**
- Pick any and evaluate it in parallel, distributed, big volumes, map-reduce, new hardware ...
### Context
Finding useful patterns in large datasets has attracted considerable interest recently,

### Problem
and one of the most widely studied problems in this area is the identification of clusters, or densely populated regions, in a multi-dimensional dataset. Prior work does not adequately address the problem of large datasets and minimization of I/O costs.

### Solution
This paper presents a data clustering method named BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies), and demonstrates that it is especially suitable for very large databases.

### (Details)
BIRCH incrementally and dynamically clusters incoming multi-dimensional metric data points to try to produce the best quality clustering with the available resources (i.e., available memory and time constraints). BIRCH can typically find a good clustering with a single scan of the data, and improve the quality further with a few additional scans. BIRCH is also the first clustering algorithm proposed in the database area to handle “noise” (data points that are not part of the underlying pattern) effectively.

### Evaluation
We evaluate BIRCH’s time/space efficiency, data input order sensitivity, and clustering quality through several experiments. We also present a performance comparison of BIRCH versus CLARANS, a clustering method proposed recently for large datasets, and show that BIRCH is consistently superior.
Connect the dots: *not* easy

Context & Related

Problem

Solution 1

Solution 2

Evaluation

Title: keyword1 keyword2

Abstract: line1 line2 line3

Intro: par1 par2 par3 par4

Conclusion: par1 par2 par3
In Databases

select
from
where

Joke that only works in Portuguese:
the word for data (dados) is the same for dice (dados)
Context and Problem

The state of the art
The state of the art

System

Producer

Consumer

overlay network (brokers)

XML???

REAL EXAMPLE

input

Routing table

Message Filtering

queries

results

queries

results
Problem and Goal

Producer
- XML

Consumer
- query
- query
- query

System
- Routing table
- input
- Message Filtering

Overlay Network (brokers)
- queries
- results

REAL EXAMPLE
Evaluation

Find out where and why!
My PhD has almost died right here

- No f*ing pattern in the results!!!!!
Look at the big picture & Play with your variables (data + queries)

Moro et al @ VLDB 2005
My (& students’) Work

Listar

- Eventos - Periódicos - Trabalhos Recentes

Curadoria

- Adicionar Evento - Inserir Trabalhos em Eventos - Fluxograma de Atividades

Destaques

- A BDBComp possui atualmente 14134 trabalhos publicados em periódicos nacionais e anais de eventos realizados no Brasil.
- Use o serviço “Buscar na Web” sempre que o texto completo de algum trabalho catalogado na BDBComp não estiver disponível.
- A BDBComp inclui também trabalhos publicados nos seguintes periódicos: JBCS, RITA, IP, INFOCOMP, RBIE e RB-RESD.

VIANA, W. & MORO, M. M. Busca de Caminhos entre Usuários de Redes Sociais em Tempo Real. BRASNA M 2012
Context

40% socialize more online than they do face-to-face.

Every minute of the day:

- 684,478 pieces of content are shared on Facebook
- 3,600 photos are shared on Instagram
- 100,000 tweets are sent
- 2 Million queries are searched on Google
- 48 hrs of video are uploaded to YouTube

Some statistics:

- 1 Billion # REGISTERED USERS
  - FACEBOOK: 552 Mil
  - TWITTER: 6.9 Mil
  - GOOGLE+: 75 Mil
- 517 Mil DAILY ACTIVE USERS
- 400 Mil

Mirella M. Moro
Question
who are those people?
Solution: the traditional one
Solution: the traditional one

huuuuuuge

Not accessible
Solution: the traditional one
Problem

1. Without building whole network
2. Confined to a small part
3. In real time

• Short path
Solution

Search
Bidirectional Search
Excellent idea

Not enough
Clustering
Geo-located
Bidirectional search
+ (geo) directed
Friendrouter

Find how you connect. To anyone.

The theory of 6 degrees of separations says at most 6 people separate you from anyone else in the world. Now you can find out who those people are. Friendrouter will help you discover how your favorite Twitter personality is connected to you.

Sign in with Twitter
Friendrouter
Find how you connect. To anyone.

Start at: @
Provide the Twitter user you wish to route to you.

End at: @mirellammore
Twitter user where the route ends (you).

Filters: 
Optional, will not include users specified here on the route.

Post a tweet on my timeline with the route.

Find Route

Your last routes
{luisvonahn}
{laryo}
{marieforgue}
{jeffjarvis}
{reaimilakunis}
{c_ashtonkutcher}

Find a route | About | Contact
mirellammmoro > taylorswift13

Friendrouter

Find how you connect. To anyone.

8 nodes explored. Target location: Belo Horizonte, MG, Brazil.
Obtaining profile info...

taylorswift13  taylornation13  portalswift  sessaoextra

(portalswift  sessaoextra)

(NEEDTObREATHE)
Route found!
5 people in between.

taylorswift13  taylornation13  portalswift  sessaoextra  Analice  guhcampus  mirellammoro

http://friendrouter.com/route/51e6fb6ebc2e8b75b789b6bc
Route found!
6 people in between.
22 nodes, 13 for goalset. Total: 35

BahaAlimov  teemonya  RealMiguelCotto  juanes  rubiapria  Mobilon  PedroVillalobos  wladston
<table>
<thead>
<tr>
<th>Caminho</th>
<th>Vértices explorados</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>partindo do alvo</td>
</tr>
<tr>
<td>A</td>
<td>8</td>
</tr>
<tr>
<td>B</td>
<td>17</td>
</tr>
<tr>
<td>C</td>
<td>18</td>
</tr>
<tr>
<td>D</td>
<td>4</td>
</tr>
<tr>
<td>E</td>
<td>12</td>
</tr>
<tr>
<td>F</td>
<td>11</td>
</tr>
<tr>
<td>G</td>
<td>2</td>
</tr>
<tr>
<td>H</td>
<td>2</td>
</tr>
</tbody>
</table>
And Beyond
Grad - MSc - PhD
Why?

“Feeling” about student’s paycheck…

Time

Work in industry

Time

Grad School

$BSc$

$BSc$

$MSc$

$PhD$

$\text{Mirella M. Moro}$
Average salary
Programmers
Brazil. SouthEast
[SP-RJ-ES-MG]
Watch out!
Grad School is Career
Grad School = Reading + Analyzing
Grad School Alone + in Groups*
specially in Databases

* Group = friends @ face/G+, classmates, professors, including from other areas
Grad School is Scientific Events

Work + Travel → Scientific Event
Grad is Hard

“No country for old men”
You

Knowledge

Vision

Studying

Experience

Solving Problems

You will never, ever, regret spending time on yourself & your studies
3ª Escola de Verão em Computação

Belo Horizonte
3 a 7 de Fevereiro de 2014
http://www.evcomp.dcc.ufmg.br
evcomp@dcc.ufmg.br

Realização

Apoio

www.dcc.ufmg.br/~mirella