Big Data Visual Analytics

Jean-Daniel Fekete
AVIZ/INRIA
Jean-Daniel.Fekete@inria.fr
http://www.aviz.fr/~fekete
Big Data

• Volume
  – Like “really big”, has evolved with time from Tb to Pb

• Variety
  – Many types, e.g. text, image, tables

• Velocity
  – Acquisition/input speed, output speed

• Variability, Veracity...Vatever

• Traditionally used with predictive analytics
The Visual Analytics Process

Exploration and Latency

3 types of latency to consider for HCI:

1. **Continuity Preserving Latency**: $\sim 0.1s$ user feel that the system is reacting instantaneously

2. **Flow Preserving Latency**: $\sim 1s$ user’s flow of thought to stay uninterrupted

3. **Attention Preserving Latency**: $\sim 10s$ keeping the user’s attention focused on the dialogue


Scaling Visualization

• Vis. does not scale well
  – Not in number of items
  – Not in number of dimensions

• It needs additional methods such as:
  – Sampling (of items/dim.)
  – Aggregation
  – Dimensionality Reduction

• These methods introduce artifacts
  – Their results should be explored too, to be validated!
Layers of Visual Analytics

Three Layers:

• Data Management
• Analytics
• Visualization+ Interaction
Examples

- Hierarchical Clustering Explorer
- WikiReactive
- HAL Deduplication Framework
- Real-time sentiment analysis
- Nanocubes
- Progressive tSNE
Hierarchical Clustering Explorer  (Seoh & Shneiderman 2002)

http://www.cs.umd.edu/hcil/hce/

• Data
  – Multidimensional (n numerical dimensions)

• Task
  – Find clusters that clearly reflect properties in the data

• Volume: In memory

• Variety: none

• Velocity: none
Hierarchical Clustering

Initial Data Items

Distance Matrix

<table>
<thead>
<tr>
<th>Dist</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>20</td>
<td>7</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td></td>
<td>10</td>
<td>25</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td></td>
<td></td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>D</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Hierarchical Clustering

Initial Data Items

Distance Matrix

<table>
<thead>
<tr>
<th>Dist</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>20</td>
<td>7</td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>B</td>
<td>10</td>
<td>25</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td></td>
<td></td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>D</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Hierarchical Clustering

Single Linkage

Current Clusters

Distance Matrix

<table>
<thead>
<tr>
<th>Dist</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>20</td>
<td>7</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>10</td>
<td>25</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td></td>
<td></td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>D</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Hierarchical Clustering

Single Linkage

Current Clusters

Distance Matrix

<table>
<thead>
<tr>
<th>Dist</th>
<th>AD</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>AD</td>
<td>20</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td></td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Hierarchical Clustering

Single Linkage

Current Clusters

Distance Matrix

<table>
<thead>
<tr>
<th>Dist</th>
<th>AD</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>AD</td>
<td>20</td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>B</td>
<td></td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Hierarchical Clustering

*Single Linkage*

Current Clusters

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>D</th>
<th>C</th>
<th>B</th>
</tr>
</thead>
</table>

Distance Matrix

<table>
<thead>
<tr>
<th>Dist</th>
<th>AD</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>AD</td>
<td>20</td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>B</td>
<td></td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Hierarchical Clustering

Single Linkage

Current Clusters

Distance Matrix

<table>
<thead>
<tr>
<th>Dist</th>
<th>ADC</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADC</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Hierarchical Clustering

*Single Linkage*

**Current Clusters**

- A
- D
- C
- B

**Distance Matrix**

<table>
<thead>
<tr>
<th>Dist</th>
<th>ADC</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADC</td>
<td></td>
<td>10</td>
</tr>
<tr>
<td>B</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Hierarchical Clustering

Single Linkage

Current Clusters

A

D

C

B

Distance Matrix

<table>
<thead>
<tr>
<th>Dist</th>
<th>ADC</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADC</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Hierarchical Clustering

*Single Linkage*

Final Result

Distance Matrix

<table>
<thead>
<tr>
<th>Dist</th>
<th>ADC</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADC</td>
<td>B</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Hierarchical Clustering Explorer

http://www.cs.umd.edu/hcil/hcexplorer/
Hierarchical Clustering and Interaction

- **How many cluster?**
  - Many criteria

- **Explore interactively**
  - Vary height (distance)
  - Vary number of clusters
  - Vary distance function

- **Are they good in the end**
  - Many way to assess but linear!
What if the number of vectors increases?

- From 1,600 to 10,000?
  - ~100,000,000 entries for the distance matrix
  - Memory and Computation still OK
- From 10,000 to 100,000,000?
  - Memory and Computation not OK
What if the number of vectors increases?

- 100,000,000 vectors could fit in memory?
- The distance matrix cannot fit in memory
  - It will take hours to compute
- Interaction is not possible any more

What can we do about it?
Strategies to cope with Big Data and Visual Analytics

1) Increase the memory?
2) Use a distributed systems?
3) Use a parallel system (HPC)?
4) Use tricks?
Increase the memory for Big Data Visual Analytics

• How much?
  – Say for n = 1,000,000 (10^6)

• Dataset + distance matrix + hierarchy

• Memory = ???

• Time to compute the distance matrix?
  – Assume 10^6 operations per second

• Conclusion?
Use a distributed system

- How many machines to perform the computation quickly?
  - Say 10s

- Distributed system have a high latency
  - Usually > 10s, around 30s to minutes

- Not good for interaction

- But can compute results ahead of time
Use a Parallel System (HPC)

• Much more expensive than a distributed system
  – But faster

• Do you really need a special architecture?
Use Tricks: Hybrid Algo.

- Clustering a huge dataset?
- HC is quadratic: not possible
- K-Means is linear but requires a good K
- Sample -> HC -> Estimate good K -> k-Means
- Need a good sampling


Does not work well for Text mining
Big Data Visual Analytics

3 situations according to Hardley Wickham
https://peadarcoyle.wordpress.com/2015/08/02/interview-with-a-data-scientist-hadley-wickham/

When data does not fit in memory (1TB):

1) Data can be filtered/selected to become small
   - actually small data problems, once you have the right subset/sample/summary

2) Analysis can be split into independent chunks
   - actually lots and lots of small data problems

3) Don't know how to filter/split, hard case!
   - irretrievably big
   - Research is working on it
WikiReactive


• Collect wikipedia changes and computes derived information
  – Diffs, user contributions, user per character

Collecting recent changes

WikiReactive Database

Query recent changes

Store latest status

Query data

WikiReactive Database

Web service

Compute and maintain aggregated data

The Beatles

This article is about the band. For their self-titled album also known as The White Album, see The Beatles (album).

The Beatles were an English musical group from Liverpool whose members were John Lennon, Paul McCartney, George Harrison, and Ringo Starr. They are one of the most commercially successful and critically acclaimed bands in the history of popular music.

The Beatles are ranked as the world's best-selling music act of all time in the United States, the world, and the United Kingdom. The Beatles released more than 40 different singles, albums, and EPs that reached number one in many other countries, including the United Kingdom, where they were number one a total of 29 times. In 2004, Rolling Stone magazine ranked The Beatles #1 on its list of 100 Greatest Artists of All Time, their innovative music and cultural impact helping define the 1960s, and their influence today.

The Beatles led the mid-1960s musical "British Invasion" into the United States. Although their initial rock and roll and homegrown style, the group explored genres ranging from Tin Pan Alley to psychedelic rock
classical (even with the inclusion of elements from the Beatles's early days in Liverpool). They were skilled musicians, songwriters, and performers, and their songs and performances were a primary influence on the evolution of popular music.
WikiReactive

• Volume
  – 5 million articles in English, many TB of text

• Variety
  – Text + previous versions, structure
  – Users (id), Talks, categories, stats

• Velocity
  – About 100 changes per second
  – But each article does not change every second

• HW Category (1, 2, or 3)?
HAL Deduplication framework

• For each article author added to the HAL database

  Hal deduplication workflow

  - XML parser
  - Fill database
  - Metric calculation
  - Or split
  - Or join
  - Visualization module
  - Decision-making module

• Computes similarity with all other authors
• Resolve simple case (< or > threshold)
• Show an interface for the other cases
HAL Deduplication framework

• Volume
  – 3 million articles, many TB of text, 3 million authors

• Variety
  – Users (id, email, institution, lab, date)

• Velocity
  – About 1 change per second

• HW Category (1, 2, or 3)?
Real-Time Sentiment Analysis

- For each new document scrapped
  - Compute part-of-speech tagging, lemmatization, negation detection, feature extraction, sentiment detection, sentiment-to-feature mapping

"password" (overall 129 comments):

I usually order online. This time my password was not accepted. I had to complete order by phone. The MFR Spoke to was very helpful. She was also very friendly.

Online website is confusing. Had difficulty with password. Called number on screen and got knowledgeable caller for who closed the call.
Real-Time Sentiment Analysis

- **Volume**
  - many million articles read continuously
- **Variety**
  - Time-stamp, text
- **Velocity**
  - As the crawler can work
- **HW Category (1, 2, or 3)?**
Nanocubes (Lins et al. 2013)

Nanocubes

- Create a spatio-temporal index
- Quickly retrieve distributions from range-queries
  - Over time
  - Over space
  - Over values
- Index creation can take hours
Nanocubes

- Volume
  - Many (200) million points
- Variety
  - Spatio-temporal data
- Velocity
  - Static
- HW Category (1, 2, or 3)?
Beyond Pre-Computation: Bounding Time and Quality

• Visualization is User Centric
  – Visualization will only show a small amount of data
  – Visualization need interactive time
  – How can we address the scale in interactive time?

• Analysis is Program Centric
  – Analysis will read data, process it and store its results in the end
  – Analysis will produce unbounded amounts of data in unbounded time
  – How can we get something in a bounded time?

• Databases is Data Centric
  – Databases will store and retrieve unbounded amounts of data in unbounded (but fast) time
  – How can we bound time with a specified level of quality?
Progressive VA

• Allow Exploratory tools to work while the computation is being done


Progressive MDS
Progressive tSNE (Pezzotti et al. 2016)

- Multidimensional projection method
- Input: points in nD
- Output: points in 2D
- Similar points nearby

https://lvdmaaten.github.io/tsne/
Progressive tSNE

- Compute distances
- Iterate to converge
Progressive tSNE

• Volume
  – many million points
• Variety
  – N-D points
• Velocity
  – Static or dynamic
• HW Category (1, 2, or 3)?
Bibliography


