BigDataCHALLENGE

TELECOM ITALIA

«VISUALIZATION OF POTENTIAL CUSTOMERS»

Cubas Saiz, Tinguaro.

Pérez Bello, Miguel.

Rodríguez Pardo, Guillermo.

Team: ETSII ULL

Motivations

We love to innovate in developing software and this contest gives us the opportunity to learning more about technologies that we have never used until now like Unity3D, R to analyze data, etc.

Our main motivation at the moment of developing the proposal is to give a way of getting better benefit in business models, offering the most expensive and top quality products to the customer with more economical resources and that presents a major consume in calls, sms and internet.

Goal

The main goal is to design a way to focus the products sales, after data analysis, giving to the application users a “nice and friendly visualization” that allows them to achieve conclusions about data.

We include in visualization many tools developed to manipulate the data views, so we can have more precision to offer some products.

Idea

"Focusing the sales of Telecom products/fares to the most potential customers, after using Clustering techniques over a data set geolocated in Milano, giving a new type of geographic visualization ."

i. We have taken the provided geolocated grid of Milano.

ii. We have characterized Telecom fares in 5 types (we made a cluster per fare/offer).

*Observation: Quantity / Price ('---' means not contemplated in offer).

	Calls	SMS	Internet	Other Servicies
Product-Fare 1	LOW / LOW	LOW / LOW	LOW / LOW	--- / ---
Product-Fare2	LOW / LOW	HIGH / HIGH	LOW / LOW	--- / ---
Product-Fare 3	HIGH / HIGH	LOW / LOW	HIGH / HIGH	--- / ---
Product-Fare 4	HIGH / HIGH	HIGH / HIGH	HIGH / HIGH	--- / ---
Product-Fare 5	HIGH / HIGH	HIGH / HIGH	HIGH / HIGH	HIGH / HIGH

Idea

iii. We have **aggregated cells** data by calls, sms and Internet traffic (considering only the country code of Italy, 39) and we have **applied the K-Means algorithm** for clustering.

iv. We have **linked characteristics of each fare to each cluster** based in a scale from minimum to maximum price/consume/service quality (to characterize the finals clusters).

v. With the obtained clusters, we have **analyzed the reliability** of the predictions made with the distances between each element of a cluster and the cluster average.

vi. Finally, we have **used Unity3D to design** a standalone application with **a new way of data visualization** in which we include several information like final clustering, reliability of product offered (standard deviation with respect to cluster average) , and specific information by cells like internet traffic and calls and sms density categorized on ‘low’, ‘medium’, ‘high’, etc.

Technical description

- More technical details are annexes in this template but we have to comment the following:

- **Use of K-Means algorithm** to clustering the cells of the Milano's grid with the cluster (fare / product) with minor distance.
- Distance equation: **extrapolated Euclidean distance.** (Note: 'i' goes from 1 to 10.000 cells and 'j' goes from 1 to 5).

$$Distance\left(Cell_i, Cluster_j\right) = \sqrt{\left(Cell_i.X - Cluster_j.X\right)^2 + \left(Cell_i.Y - Cluster_j.Y\right)^2 + \left(Cell_i.Z - Cluster_j.Z\right)^2}$$

- Where:
  - **X**: is the **calls** density.
  - **Y**: represents the **SMS** density.
  - **Z**: represents the **Internet** traffic.

- Averages to redistribute clusters' centroids and to get the behavior that K-Means provides.

Used data

- We have focused our analysis in a Dataset of Milano called **“Telecommunications – SMS, Calls, Internet – MI”:**

  - This dataset provides information each 10 minutes about calls, sms and Internet activity in the metropolitan area of Milano at November and December.

  - **Data used**: over 320.000.000 registers (160.000.000 per month) were analyzed with ‘R’ (statistical Software) on a private server of 64 Gb (RAM).

Used data

- Attributes used:

  - **square_id:** to characterize each region.
  - **sms_out**: to calculate the traffic of out sms.
  - **call_out**: to get a factor for measure the call level of each region.
  - **internet_traffic**: to evaluate the use of Internet in each region.
  - **country_code**: for filtering data.

- **Filter**: we have only used sms and calls with “country_code = 39” (from Italy).

Main outcomes

- The outcomes will be explained in two parts: data analytics and visualization.

We have analyzed a dataset about Milano's population, and our clustering's algorithm has discovered an economical ratio, used to obtain the relationship between a set of offers distributed from min. to max. price and the cells geolocated in Milano (provided by the API).

The most important outcome is the innovation that supposes using **Unity3D** (created to design games) to obtain many ways of visualizing dataset previously analyzed, in order to get an easy mode for getting conclusions about them.

Impact

There are many techniques for **clustering or market segmentation** that offers a lot of advantages to business models, **increasing their benefits**.

Offering a better relationship between price and quality, **adjusting the type** of the products offered to the most potential customers, is the basis for getting a good user satisfaction; also, it **becomes a better service** for customer support.

Future evolution

I. Use an external statistical dataset (for instance, from ISTAT) to obtain rent factors of Milano in order to increase the reliability of the predictions from the clustering algorithm.

II. Use another dataset, like weather or traffic, for making predictions to know when is the best moment in which potentials customers will have a great response or availability.

The proponent

Team of three undergraduate students on Computer Science from the University of La Laguna (ETSII), Tenerife (Canary Island), Spain.

We have decided to participate in this contest because of a teacher have explained us the rules and motivations to get involved.

Our interest is focused in order to get as knowledge as we can about «**mining data world**» and the experiences that the participation in a contest of this scale could give us.

Annexes

Clustering + Visualization

Link to video:

https://www.dropbox.com/sh/10jg951zkwdeqj3/3DgQVS

TUnO/VideoDemonstration#lh:null-BigDataChallenge.mp4

+ Clustering Algorithm: << K-Means >>

For each cell in Milano's grid:

Step 0 – Assign the initials clusters centroids.

Step 1 – Calculate the nearest cluster centroid.

Step 2 – For each cluster:

- Recalculate clusters centroids.

Step 3 – Repeat steps 1 and 2 while there are changes.

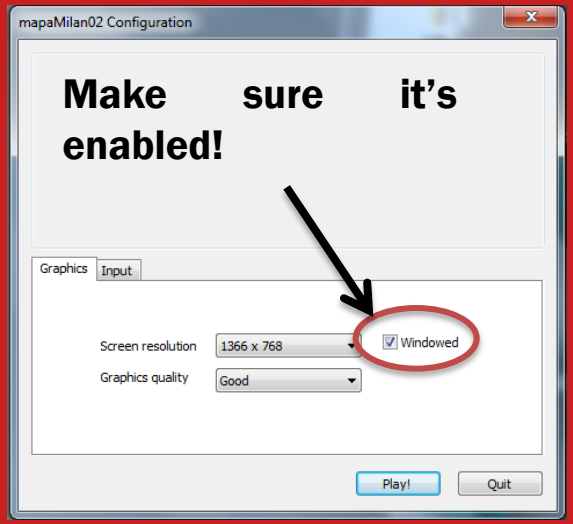
+ Relation Clusters - Fares / Telecom Products

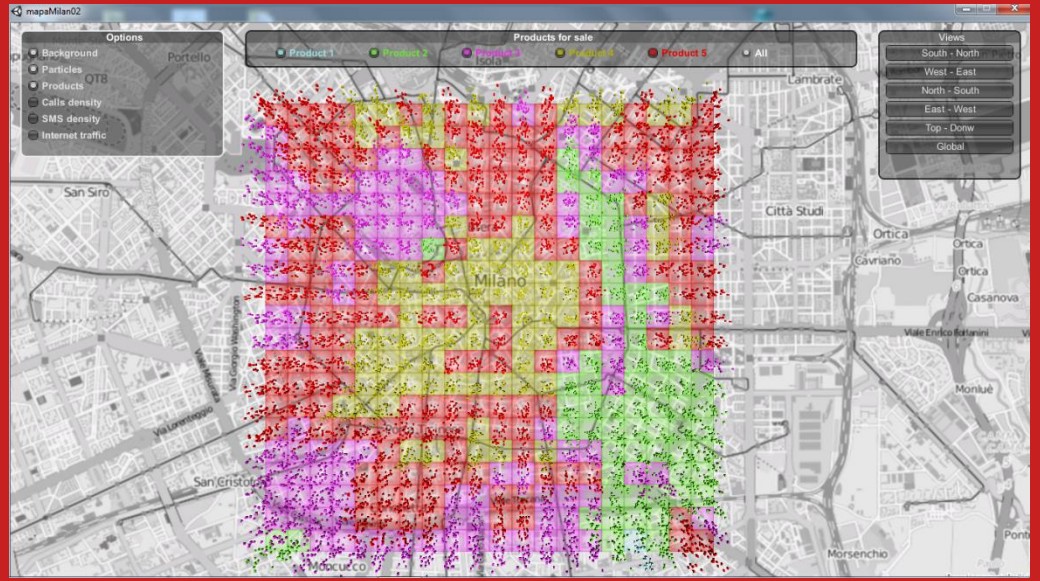
Each cluster **was created thinking in a set of fares / products** that it could be adjusted to Telecom products characteristics.

We can see the differences between the 5 clusters attending to an **economical scale** (extrapolated in a data usability scale by calls, sms and Internet).

+Prototype. GUI:

We have <u>developed a stand-alone application</u> in which we can show a few better our idea in a better way, so our results are seen more straightforward.





+Prototype. Windows:

<u>Window products</u>: To select the product / fare we want to see in Milano's grid.

<u>Window Options</u>:

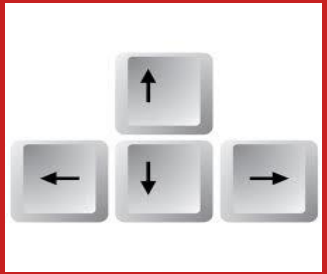
- **We can see cells background and/or cells particles.**
- **We can see the calls / sms / Internet proportion in each cell or cells group (20 x 20).**

+Prototype. Views:

Window Views: To select the way to seeing (the view perspective) the different products / fares geolocated in Milano's grid.

+Prototype. Camera Controllers:

Using arrow keys we have control over 20 x 20 Milano's grid subset and camera movement (camera follows it).

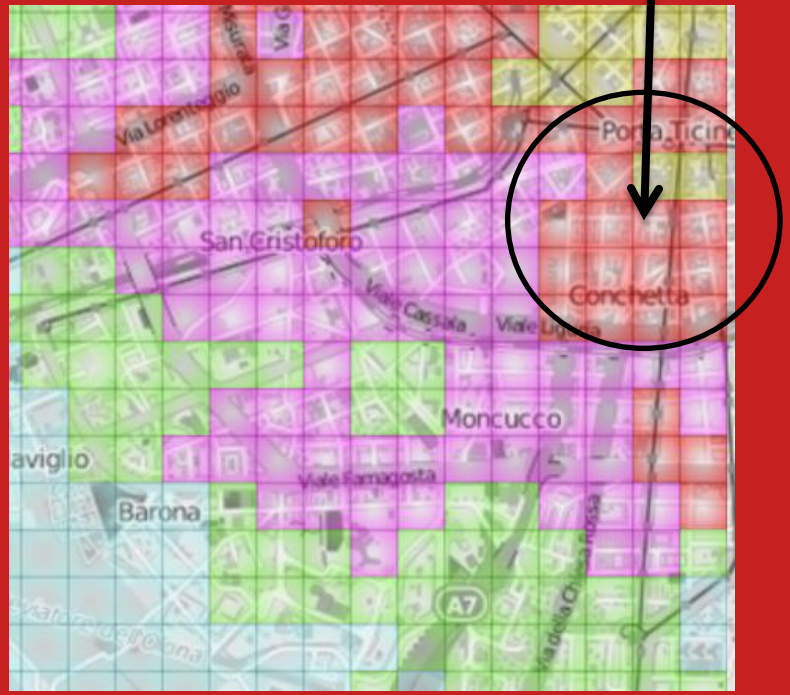


+Use Case (Examples for increasing benefits):

To put billboards advertising the most recommended product/fare on this collapsed road of calls/SMS/Internet.

+Use Case (Examples for increasing benefits):

To focalize the Telecom fares sales emission on TV or sales via calls or via flyers on populated areas like this.



BigDataCHALLENGE

TELECOM ITALIA

«VISUALIZATION OF POTENTIAL CUSTOMERS»

Cubas Saiz, Tinguaro.

Pérez Bello, Miguel.

Rodríguez Pardo, Guillermo.

ETSII ULL