INFLU-VENN-ZA

GIULIO BAMBINI – VISUAL ANALYTICS[1]

02/06/2016

CONTENTS

ABSTRACT

Il seguente elaborato si propone di presentare un'interfaccia visiva per la risoluzione di un problema analitico relativo all'esplorazione di dati scientifici. Per l'esplorazione spaziale sono stati analizzati dati riguardanti relazioni tra diversi ceppi influenzali che affliggono il genere umano e diverse categorie animali. Per l'esplorazione temporale sono stati analizzati dati riguardanti le serie storiche per ciascun ceppo influenzale distribuite dal 1917 al 2015.

[1] *Corso di laurea magistrale in Business Informatics, Università di Pisa*

1 INTRODUZIONE

Influ-Venn-Za è una visualizzazione web interattiva che permette l'esplorazione di diverse metriche di analisi:

1. Relazioni tra ceppi influenzali e generi
2. Distribuzione temporale dei ceppi influenzali
3. Distribuzione spaziale dei ceppi influenzali

L'idea nasce dall'infografica *Influ-Venn-Za - Who can catch this flu?* pubblicata sul sito web informationisbeautiful[2] e supportata da una serie di dati aggregati per i ceppi influenzali disponibili su Google Spreadsheets[3]. La scelta di puntare sulla rappresentazione insiemistica è stata dettata dalla possibilità di dare risalto ad alcune relazioni sui dati in maniera immediata, schematizzata ed altamente informativa per l'utente che ne visualizza il contenuto. Per le distribuzioni temporali e spaziali dei ceppi influenzali sono stati utilizzati dati disponibili sul sito ncbi.nlm.nih.gov[4]. Queste ulteriori aggregazioni hanno permesso di ampliare gli scopi del progetto iniziale aggiungendo ulteriori rappresentazioni grafiche per evidenziare e profilare maggiormente i dati analizzati nella rappresentazione insiemistica.

2 STATO DELL'ARTE

Sul web sono presenti diversi progetti che trattano il tema della diffusione epidemica dell'influenza. Il modello di visualizzazione comunemente utilizzato consiste nella rappresentazione di una mappa di calore o di una mappa con i marker posizionati geograficamente, dove ogni marker corrisponde ad un ceppo influenzale.

name	url
cdc.gov	http://www.cdc.gov/flu/weekly/usmap.htm
everydayhealth	http://www.everydayhealth.com/flu/map/
flunearyou	https://flunearyou.org/
healthmap	http://vaccine.healthmap.org/
webmd	http://symptoms.webmd.com/cold-and-flu-map-tool/

Tabella1: Lista dei servizi presenti sul web.

Le informazioni dei dati sulla mappa vengono messe in risalto attraverso la misura di intensità del colore di una zona geografica che definisce solitamente più fattori di analisi; ad esempio la gravità di diffusione di un tipo di influenza o la pericolosità di infezione. Le limitazioni evidenziate da questo genere di visualizzazione sono l'assenza di analisi sulle relazioni tra i tipi di influenza e l'assenza di un'esplorazione della dimensione temporale. Nessun servizio citato mette a disposizione dell'utente un quadro chiaro di alcune misurazioni statistiche sui dati. Per esempio non si è in grado di capire ne' il picco temporale di diffusione ne' le occorrenze esatte per quanto riguarda la distribuzione dell'influenza per paese o totale. Non esiste inoltre sul web una visualizzazione insiemistica interattiva delle varianti influenzali.

[2] *http://www.informationisbeautiful.net/visualizations/which-flu-virus/*
[3] *https://goo.gl/ggWgc6*
[4] *http://www.ncbi.nlm.nih.gov/genomes/FLU/FLU.html*

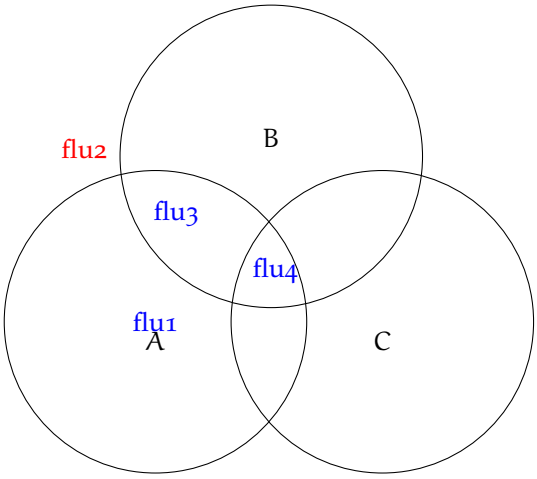
3 RAPPRESENTAZIONE E ANALISI DEI DATI

Per rappresentare in una forma corretta le relazioni tra i dati aggregati sono state scelte quattro tipologie di visualizzazione, il diagramma di Eulero-Venn, il bar chart, il line chart e il pie chart per la rappresentazione spaziale e temporale dei dati.

3.1 Diagramma di Eulero-Venn

Il diagramma di Eulero-Venn viene normalmente utilizzato per trasmettere i concetti di insieme, elementi e relazione tra elementi. Per gli insiemi con un numero finito di elementi è possibile rappresentare ciascun insieme come una figura chiusa delimitata da una linea, mentre nel caos dei ceppi influenzali ciascun elemento verrebbe indicato come nella rappresentazione sottostante:

$E = \{flu1, flu2, flu3, flu4\}$



Per ciascun elemento si presenta una relazione differente:

- $flu1 \in A - (B \cup C)$
- $flu2 \in (A \cup B \cup C)^C$
- $flu3 \in A \cap B$
- $flu4 \in A \cap B \cap C$

Tali relazioni tra insiemi ed elementi possono essere riproposte sotto forma di dati strutturati. A ciascun elemento vengono assegnati degli attributi che ne definiscono le proprietà. Per esempio è possibile recuperare i valori delle coordinate assegnate agli elementi per mapparne la posizione e i valori degli insiemi in cui è incluso ciascun elemento.

```
#!/Influ-Venn-Za/script/example_1.json
{
  "name" : "flu1",
  "canInfect" :
  [
    "A",
  ],
```

```
  "coordinates":
  {
    "x": 1.2,
    "y": 1.2
  }
},
{
  "name" : "flu2",
  "canInfect" :
  [
    "null",
  ],
  "coordinates":
  {
    "x": 2.0,
    "y": 2.0
  }
},
{
  "name" : "flu3",
  "canInfect" :
  [
    "A",
    "B",
  ],
  "coordinates":
  {
    "x": 0.7,
    "y": 0.7
  }
},
{
  "name" : "flu4",
  "canInfect" :
  [
    "A",
    "B",
    "C",
  ],
  "coordinates":
  {
    "x": 0.33,
    "y": 0.35
  }
},
```

Listing 1: Esempio di dati strutturati per Diagramma di Eulero-Venn.

Lo stesso procedimento è stato applicato per ciascun elemento g creato con D3 prendendo i dati da data.json. Nel caso di Influ-Venn-Za vengono creati 25 elementi text all'interno di 25 elementi g corrispondenti a differenti varianti di ceppi influenzali.

```
#!/Influ-Venn-Za/script/example_2.js

// Aggiunge flu_variants
svg_canvas.selectAll("g.item")
  .data(flu_variants)
    .enter()
    .append("g")
    .attr("class", "item")
    .append("text")
    .attr("x", function(d) { return d.coordinates.x })
    .attr("y", function(d) { return d.coordinates.y })
    .attr("fill", "#FFF")
```

```
    .attr("font-weight", "bold")
```

Listing 2: Esempio di risultante elementi g da D3 a HTML.

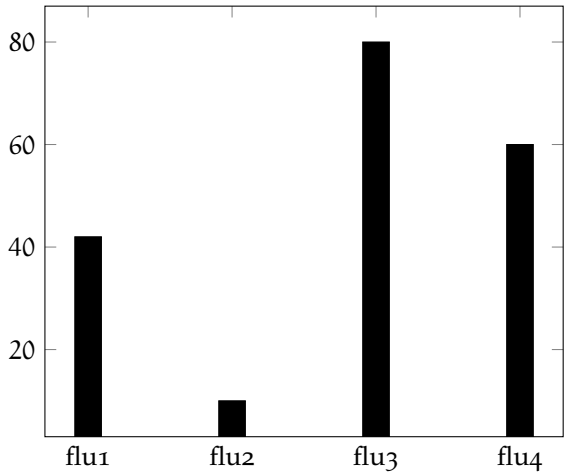
```
#!/Influ-Venn-Za/script/example_2.html

//esempio di risultante in HTML per un solo oggetto g
<g class="item">
  <text x="675" y="310" fill="#FFF" font-weight="bold">H1N2</text>
</g>
```

Listing 3: Esempio HTML generato da D3.

3.2 Bar chart, line chart e pie chart

Il bar chart è un grafico che rappresenta dati aggregati attraverso rettangoli di lunghezza proporzionale ai valori che rappresenta. Nel caso dei ceppi influenzali sono state rappresentate le occorrenze totali come nell'esempio sottostante:



In questo caso i dati ricevuti dal grafico sono stati strutturati come una lista di array con ciascuno una coppia elementi, il primo per identificare la variante influenzale e il secondo per identificarne l'occorrenza.

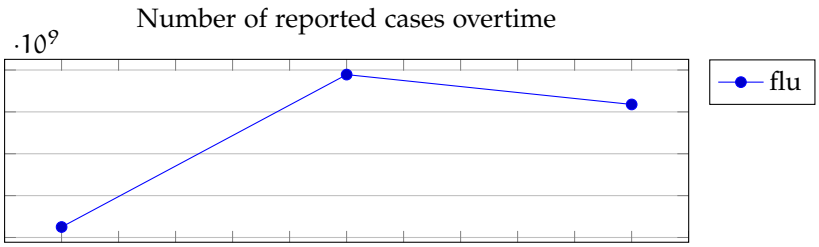
```
#!/Influ-Venn-Za/script/example_5.js

0:Array[2]
  0:"flu1"
  1:42
1:Array[2]
  0:"flu2"
  1:10
2:Array[2]
  0:"flu3"
  1:80
3:Array[2]
  0:"flu4"
  1:60
```

Listing 4: Esempio dati in input per barchart.

Il line chart è un grafico che rappresenta dati aggregati (solitamente per anno/mese) attraverso serie di datapoint connessi tra loro. Nel caso dei

ceppi influenzali sono state rappresentate le occorrenze per anno come nell'esempio sottostante:



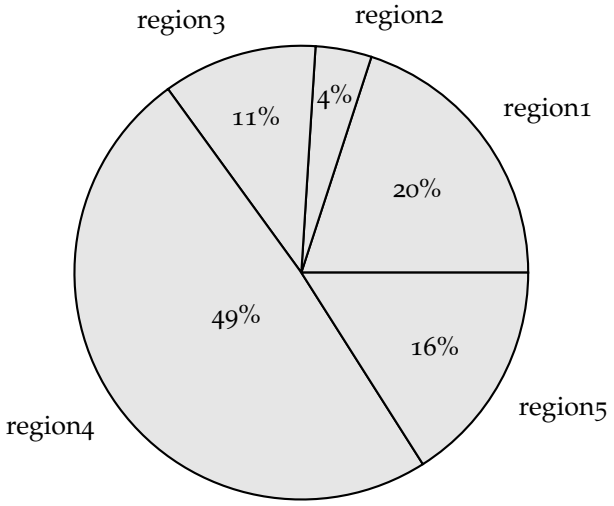
In questo caso i dati ricevuti dal grafico sono stati strutturati come una lista di oggetti composti da una coppia di elementi, il primo per identificare il timestamp a 12 cifre e il secondo per identificarne l'occorrenza associata.

```
#!/Influ-Venn-Za/script/example_6.js

0:Object
  x:368143200000
  y:1
1:Object
  x:386460000000
  y:8
2:Object
  x:454888800000
  y:5
```

Listing 5: Esempio dati in input per linechart.

Il pie chart è un grafico statistico circolare suddiviso in fette, dove la larghezza dell'arco di ciascuna fetta è proporzionale alla quantità che rappresenta. Nel caso dei ceppi influenzali sono state rappresentate le occorrenze per paese convertite in percentuale come nell'esempio sottostante:



In questo caso i dati ricevuti dal grafico sono stati strutturati come una lista di oggetti composti da una coppia di elementi, il primo per identificare l'occorrenza e il secondo per identificarne il paese.

```
#!/Influ-Venn-Za/script/example_7.js

0:Object
  count:200
  region:"Region1"
1:Object
  count:40
  region:"Region2"
2:Object
  count:110
  region:"Region3"
3:Object
  count:490
  region:"Region4"
4:Object
  count:160
  region:"Region5"
```

Listing 6: Esempio dati in input per pie chart.

4 IL PROGRAMMA

Il programma è strutturato in 8 funzioni separate. I dati sono inizialmente letti in memoria e al caricamento avvenuto viene invocata la funzione venn_vis (data) che crea la visualizzazione del diagramma di Eulero-Venn in alto a sinistra. venn_vis_data contiene a sua volta 4 funzioni. Le prime due (distance e circleCircleIntersection) calcolano rispettivamente la distanza tra i centri e i punti di intersezione di due insiemi. Le altre due funzioni (zoom e resetZoom) definiscono quale area del diagramma è stata cliccata, creano il path della forma dell'intersezione tra 2 o 3 insiemi e gestiscono l'aspetto legato allo zoom. In seguito viene chiamata la funzione fetch_and_setup_summary() dopo che i dati sono stati letti in memoria. Tale funzione legge i file in csv e crea 3 ulteriori visualizzazioni grafiche chiamando la funzione renderSummary (data), la quale ripulisce i dati e a sua volta chiama 3 funzioni di highcharts per rappresentare i grafici: summary_bar_chart (data) per il grafico a barre, linechart_vis(data) per l'istogramma, and regions_pie_chart (data) per il grafico a torta. I metodi drilldown (variant) e drilldown_host (host) aggiornano i grafici di highcharts nel momento in cui vengono selezionati sottoinsiemi dei dati, per esempio la selezione di uno specifico ceppo influenzale, o la selezione di diverse categorie di "hosts".

```
#!/Influ-Venn-Za/script/example_3.js

var raw_data;

$(document).ready(function(){
  d3.json("./json/data.json", function(error, json) {
    if (error) return console.warn(error)
    raw_data = json
    venn_vis(raw_data)
    fetch_and_setup_summary()
  })
})


var venn_vis = function(data) {...
}
```

```
var fetch_and_setup_summary = function() {...
}

var renderSummary = function() {...
}

var drilldown = function(variant) {...
}

var drilldown_host = function(host) {...
}

var summary_bar_chart = function(dataset) {...
}

var linechart_vis = function(dataset) {...
}

var regions_pie_chart = function(dataset, variant) {...
}

function selectHost(ix) {...
}
```

Listing 7: Struttura del programma.

4.1 Funzione venn_vis

La funzione viene chiamata dalla callback d3.json dopo che i dati sono stati caricati ed è responsabile della visualizzazione del diagramma di Venn e delle sue interazioni. Inoltre la funzione si interfaccia con due div, la lista degli host selezionati ( ) una volta che la visualizzazione è zoommata, e il div in alto a sinistra che restituisce le coordinate al movimento del mouse all'interno del container SVG ( ). Chiama infine i metodi drilldown e drilldown_host quando una porzione del diagramam di Venn viene cliccata.

SOTTO FUNZIONI DI VENN_VIS

La funzione circleCircleIntersection trova i punti di intersezione tra due cerchi.
L'argomento della funzione ha il centro di ciascun cerchio descritto dalle chiavi "cx" e "cy", mentre il raggio è descritto dalla chiave "radius".
La funzione zoom controlla tutti gli aspetti legati all'interazione sui cerchi del diagramma. Prende due argomenti, il primo corrisponde ai dati associati al cerchio che è stato cliccato(cx,cy,radius,fill) e il secondo corrisponde ad un eventuale indice che in questo caso non è utilizzato. Analizza la regione che è stata cliccata e calcola se i punti di selezione rientrano in uno dei cerchi. Successivamente identifica tutti gli "hosts" che rientrano dentro quella regione attiva che è stata selezionata e li inserisce un una lista definita "matching_hosts". Tale lista viene inserita sopra il barchart nel div # selected_hosts. Successivamente calcola una stringa che descrive il path della forma della regione attiva e la inserisce in una variabile chiamata "intersection_path". La regione attiva viene creata in colore giallo per evidenziarne i confini. L'intera visualizzazione del diagramma di Venn viene poi scalata e traslata per centrarla nella regione di interesse. Questa transizione impiega circa 1 secondo.
La funzione resetZoom rispristina la visualizzazione completa del diagramma

di Venn e cancella la lista dei "selected_hosts" sopra il grafico a barre.

4.2 Funzione fetch_and_setup_summary

La funzione viene chiamata dalla callback d3.json dopo che i dati sono stati caricati. Utilizza la libreria async.js per caricare la lista completa dei csv e raccogliere i loro dati. Viene creato il menu a tendina con la lista completa degli "hosts" e viene chiamata la funzione renderSummary.

4.3 Funzione renderSummary

La funzione chiamata da fetch_and_setup_summary aggrega i dati per variante influenzale e chiama la funzione di Highcharts summary_bar_chart per creare il grafico a barre in alto a sinistra. Aggrega i dati per data e chiama la funzione di Highcharts linechart_vis per creare un'istogramma in basso a sinistra. Aggrega infine i dati per paese e chiama la funzione di Highcharts regions_pie_chart per creare un grafico a torta.

4.4 Funzione drilldown

La funzione viene chiamata da venn_vis quando un elemento text per una variante influenzale viene selezionato. Come argomento di funzione ha la singola variante che le viene passata dalla selezione(come ad esempio H5N1). È responsabile del filtro sui dati e aggiorna le 3 visualizzazioni di Highcharts con i nuovi dati selezionati.

4.5 Funzione drilldown_host

La funzione viene chiamata da venn_vis quando un elemento cerchio per un singolo host specifico viene selezionato. Ha come argomento il nome dell'host che è stato cliccato. Anch'essa è responsabile del filtro sui dati e aggiorna le 3 visualizzazioni di Highcharts con i nuovi dati selezionati.

4.6 Funzioni Highcharts

Le funzioni summary_bar_chart, linechart_vis e regions_pie_chart sono 3 funzioni fornite dalla libreria Hicharts che permette di disegnare i dati secondo diversi plot.

4.7 Funzione selectHost

La funzione si occupa di gestire il controllo sull'evento onchange alla selezione di alcune regioni del diagramma di Venn. Se la selezione è avvenuta fuori da uno dei cerchi (per esempio se l'utente seleziona l'host dal menu a tendina) allora viene lanciato d3click che intercetta l'elemento selezionato attraverso dispatchEvent.

4.8 Scale

Sono state utilizzate le scale lineari con il metodo d3.scale.linear per quantificare correttamente la dimensione dei text riempiti in ciascun oggetto g rappresentativo per ciascuna variante di influenza nel diagramma di Venn.

```
#!/Influ-Venn-Za/script/example_4.js

// Ritorna il massimo(60) e il minimo(0) di humanMortalityRates
  max_hmr = _.max(
    _.map(
      flu_variants,
      function(e) { return parseInt(e.humanMortalityRate) }
    )
  )
  min_hmr = _.min(
    _.map(
      flu_variants,
      function(e) { return parseInt(e.humanMortalityRate) }
    )
  )
  font_size_scale = d3.scale.linear()
    .domain([min_hmr, max_hmr])
    .range([12, 30])

//Ritorna la scala di valori a partire da 12 fino a 30
.attr("font-size", function(d) {
      var hmr = d.humanMortalityRate == null ? 0 : d.humanMortalityRate
      var scaled_size = font_size_scale(hmr)
      return scaled_size + "px"
```

Listing 8: Esempio di struttura della scala lineare.

4.9 Librerie Utilizzate

Per la visualizzazione del summary_bar_chart, del linechart_vis e del regions_pie_chart è stata utilizzata la libreria Highcharts. Per la visualizzazione del diagramma di Eulero-Venn è stata utilizzata la libreria D3. Per la gestione e per la preparazione dei dati è stata utilizzata la liberia Underscore.js. Per la gestione delle callback è stata utilizzata la libreria async di Node.js. Per il layout responsive in HTML5 e CSS3 è stata utilizzata la libreria Bootstrap.

5 OSSERVAZIONI SUI DATI

Grazie alle metriche di analisi messe a disposizione osservando i dati sono emersi alcuni aspetti interessanti:

1. il ceppo influenzale con tasso di mortalità più elevato (H5N1) non risulta come il più diffuso globalmente;
2. i ceppi influenzali più infettivi per l'uomo sono anche i più diffusi (H1N1, H10N7, H5N1, H3N2). Esiste quindi una diretta correlazione tra diffusione globale e contagio umano;
3. i ceppi che si trovano nell'intersezione tra 3 insiemi, vedi H5N1, H7N9, H9N2, risultano meno diffusi dei ceppi che si trovano nell'intersezione tra 2 insiemi, vedi H1N1, H1N2, H3N2. Quindi non vi è una diretta correlazione tra l'aumento delle intersezioni tra insiemi e l'incremento di casi di contagio da ceppo influenzale;

4. il picco più elevato di diffusione temporale sembra configurarsi nel 2008, anno in cui si registrarono circa 8011 casi relativi al contagio da pandemia suina (H1N1). Il primo caso di contagio è stato registrato nel 1918, ciò spiegherebbe l'evoluzione del ceppo in alcune varianti similari nell'arco di 100 anni (vedi per esempio H1N2, H3N2);
5. nel periodo compreso tra il 2005 e il 2011 sembra aver avuto un impatto meno gravoso la diffusione del ceppo influenzale H5N1 responsabile del contagio da influenza aviaria. Il primo caso di contagio è datato 1977, mentre quasi il 90 percento dei paesi di diffusione è rappresentato da paesi orientali (Indonesia, Cina, Vietnam, Thailandia, Cambogia, Bangladesh);
6. gli USA risultano il paese con le percentuali di diffusione più elevate per gli insiemi di umani, foche, uccelli e maiali;
7. il Regno Unito risulta il paese con le percentuali di diffusione più elevate per l'insieme dei cavalli;
8. il Guatemala risulta il paese con le percentuali di diffusione più elevate per l'insieme dei pipistrelli;

6 CONCLUSIONI

Questo modello di visualizzazione interattiva risulta idoneo nel darci una spiegazione parzialmente corretta sull'andamento spaziale e temporale dei dati relativi alla diffusione epidemica dei ceppi influenzali e nell'offrire uno schema di visualizzazione di grosse quantità di dati comprensibile anche a non esperti del settore.