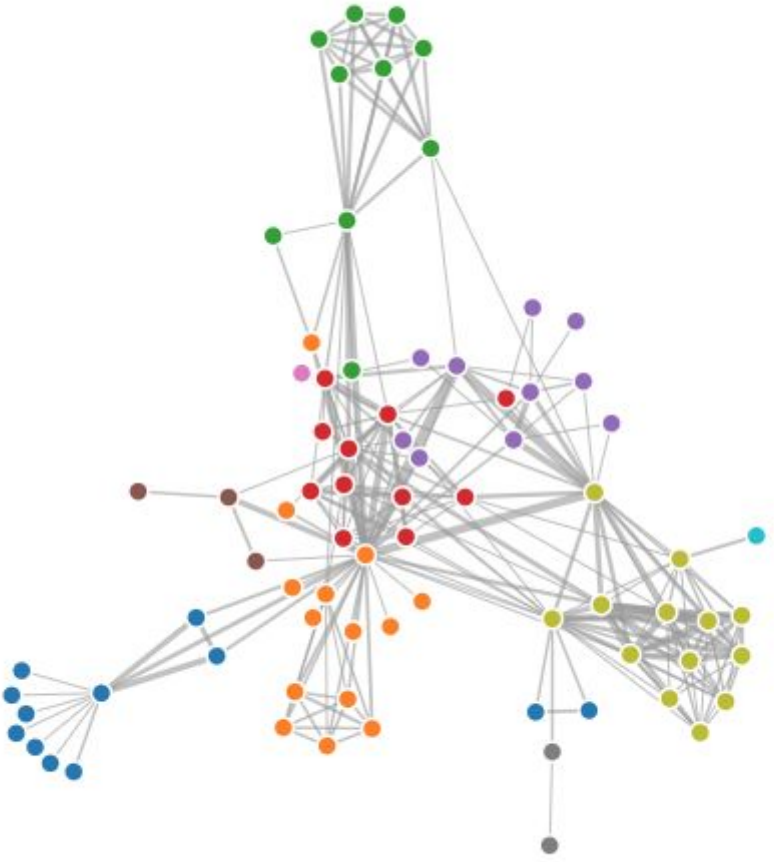
Università degli Studi di Padova
Corso di Ingegneria del Software 2020/2021



HD Viz

Visualizzazione di dati con molte dimensioni

ZUCCHETTI
IL SOFTWARE CHE CREA SUCCESSO

Oggetto dell’appalto

Le applicazioni moderne sono in grado di memorizzare e gestire volumi di dati molto elevati. Si arriva così ai Big Data, che nascono da ambiti come la telefonia, memorizzando tutte le posizioni dei cellulari in ogni momento, o il commercio elettronico con il salvataggio di tutti i click e le visualizzazioni degli utenti.
Anche i programmi tradizionali sono in grado di mantenere in linea molti più dati di quanto non avvenisse in passato, ad esempio i gestionali hanno regolarmente 10 anni di storico o la gestione delle risorse umane può arrivare a tenere 20 anni di dati per la problematica del contenzioso sul lavoro.

L’ analisi dei dati avviene sfruttando tecniche di statistica, machine learning e intelligenza artificiale. Possiamo distinguere quattro fasi fondamentali: l'acquisizione dei dati, la loro pulizia, l'analisi dei dati e infine l’interpretazione.
Nell'interpretazione dei dati distinguiamo due momenti: la prima esplorazione, chiamata EDA (Exploratory Data Analysis), e quindi la modellazione dei fenomeni trovati.

Il tipico strumento dell’EDA è la visualizzazione grafica del dato. Molto spesso un’immagine ed un grafico risultano molto più esplicativi che non una serie di numeri, infatti l’occhio umano percepisce senza difficoltà forme, gruppi e punti isolati quando i dati hanno fino a 3 dimensioni, al di sopra di questa dimensionalità l’intuizione viene a mancare.

Esempi pratici di applicazione di queste tecniche di analisi possono essere il controllo dei cedolini degli stipendi di un dipendente o le dichiarazioni dei redditi con il modello 730.

Nel caso del cedolino molti dati vengono imputati manualmente, per cui un errore di digitazione può trasformarsi in valori assolutamente inaspettati. Se 1,20 ore di straordinario diventano 120 è evidente l’effetto sul dato finale.
Ma questo controllo non può essere forzato al momento del data-entry, perché anche se palesemente improbabile è sempre un valore contenuto nel range di validità del dato.

Il modello 730 viene utilizzato annualmente per pagare le tasse. I dati sono facilmente verificabili quando provenienti dal reddito da lavoro dipendente, addirittura oggi arrivano precompilati dal ministero, ma i dati relativi alle spese detraibili e ai redditi aggiuntivi come le proprietà immobiliari, possono avere lo stesso tipo di errore che abbiamo descritto nel caso del cedolino.

In entrambe i casi ci aspettiamo che un grafico adeguato possa portare i due esempi fuori dalla distribuzione grafica della maggioranza dei dati, facendo così risaltare all’occhio umano che qualcosa non è nella normalità ma è un’anomalia da verificare.

Il presente capitolato ha per oggetto l'affidamento della fornitura per la realizzazione di un'applicazione di visualizzazione di dati con molte dimensioni a supporto della fase esplorativa dell'analisi dei dati.

Caratteristiche e Requisiti Obbligatori

L'applicazione "HD Viz" di visualizzazione dei dati a molte dimensioni sarà sviluppata prevalentemente in tecnologia HTML/CSS/JavaScript utilizzando la libreria D3.js.
La parte server di supporto alla presentazione nel browser e alle query ad un database SQL o NoSQL potrà essere sviluppata in Java con server Tomcat o in Javascript con server Node.js.

Per "molte dimensioni" si intende che i dati da visualizzare dovranno poter avere almeno fino a 15 dimensioni, deve essere possibile visualizzare anche dati con meno dimensioni.

I dati devono poter essere forniti al sistema di visualizzazione sia con query ad un database che da file in formato CSV preparati precedentemente.

"HD Viz" dovrà presentare almeno le seguenti visualizzazioni:

1. Scatter plot Matrix (fino ad un massimo di 5 dimensioni)
2. Force Field
3. Heat Map
4. Proiezione Lineare Multi Asse

La "Scatter plot Matrix" è la presentazione a riquadri disposti a matrice di tutte le combinazioni di scatter plot, con opzionalmente la distribuzione dei dati per ogni dimensione nella diagonale.
E' una delle visualizzazioni facilmente reperibili nella libreria D3.js.
Questa visualizzazione aiuta l'esploratore a trovare dimensioni con forti correlazioni e dimensioni che danno la stessa informazione.

Il grafico "Force Field" traduce le distanze nello spazio a molte dimensioni i forze di attrazione e repulsione tra i punti proiettati nello spazio bidimensionale (o anche tridimensionale).
Questo grafico esegue una riduzione dimensionale preservando, o addirittura evidenziando, le strutture presenti nei dati.

Il grafico "Heat map" trasforma la distanza tra i punti i colori più o meno intensi, facendo così capire quali oggetti sono vicini tra loro e quali sono distanti. Per una buona visualizzazione è utile accompagnare la costruzione del grafico con l'ordinamento dei dati in modo che le strutture presenti siano evidenziate, inoltre, fatta questa operazione, è possibile associare un "dendrogramma" lungo i bordi della mappa. anche questo grafico e la relativa operazione di ordinamento è facilmente reperibile tra gli esempi della libreria D3.

La "proiezione lineare multi asse" posiziona i punti dello spazio multidimensionale in un piano cartesiano, riducendo a 2 dimensioni anche dati con molte più dimensioni. Per far intuire all'analista

le strutture ed i raggruppamenti si lasciano spostare gli assi, tipicamente rendendo draggabili le frecce ad essi associate, o si crea un'animazione che ruota gli assi secondo percorsi prestabiliti. questo grafico non è tra quelli presenti negli esempi di D3, ma è visibile nel programma di data mining "Orange Canvas" o nello strumento di visualizzazione "ggobi".

"HD Viz" dovrà obbligatoriamente:

1. Ordinare i punti nel grafico "Heat map" per evidenziare i "cluster" presenti nei dati.

Requisiti Opzionali

Il tema della visualizzazione dei dati multidimensionali è vasto e ricco di spunti. Tutte le seguenti attività saranno ben accettate dal proponente:

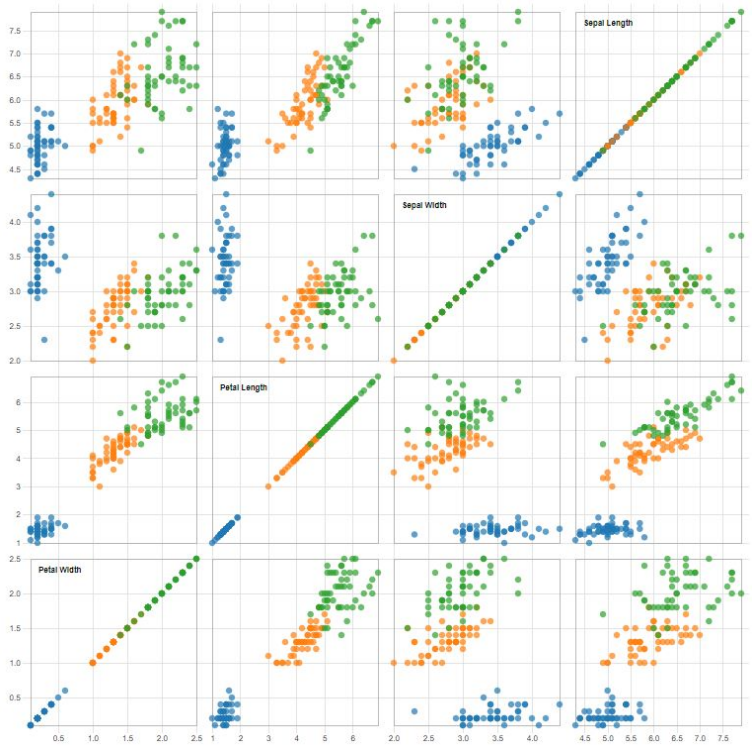
1. Altri grafici adatti alla visualizzazione dei dati con più di tre dimensioni.
2. Utilizzo di funzioni di calcolo della distanza diverse dalla distanza "Euclidea" in tutte le visualizzazioni che dipendono da tale concetto.
3. Utilizzo di funzioni di "forza" diverse da quelle previste in automatico dal grafico "force based" di D3.
4. Analisi automatiche per evidenziare situazioni di particolare interesse. Esempi di questa possibilità si possono vedere in "ggobi" e "Orange Canvas"
5. Algoritmi di preparazione del dato per la visualizzazione, cioè anziché eseguire la trasformazione direttamente nella visualizzazione far precedere un passo di trasformazione. Per questo requisito opzionale l'azienda può mettere a disposizioni librerie per gli algoritmi:
    a. t-SNE
    b. UMAP
    c. Self Organizing Map
    d. Learning Vector Quantization

Qualunque altra proposta del fornitore verrà valutata dall'azienda e eventualmente accettata come requisito opzionale se ritenuta valida per lo scopo del progetto.

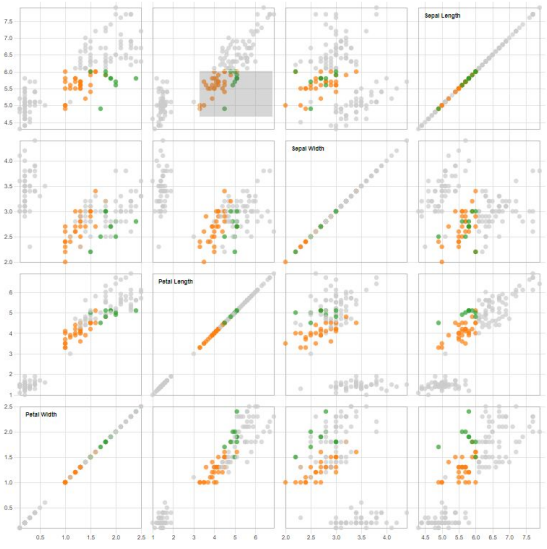
Suggerimenti

All'interno di D3 sono presenti numerosi esempi, alcuni dei quali già permettono una visione cluster dei dati. A seguire alcuni link di approfondimento:

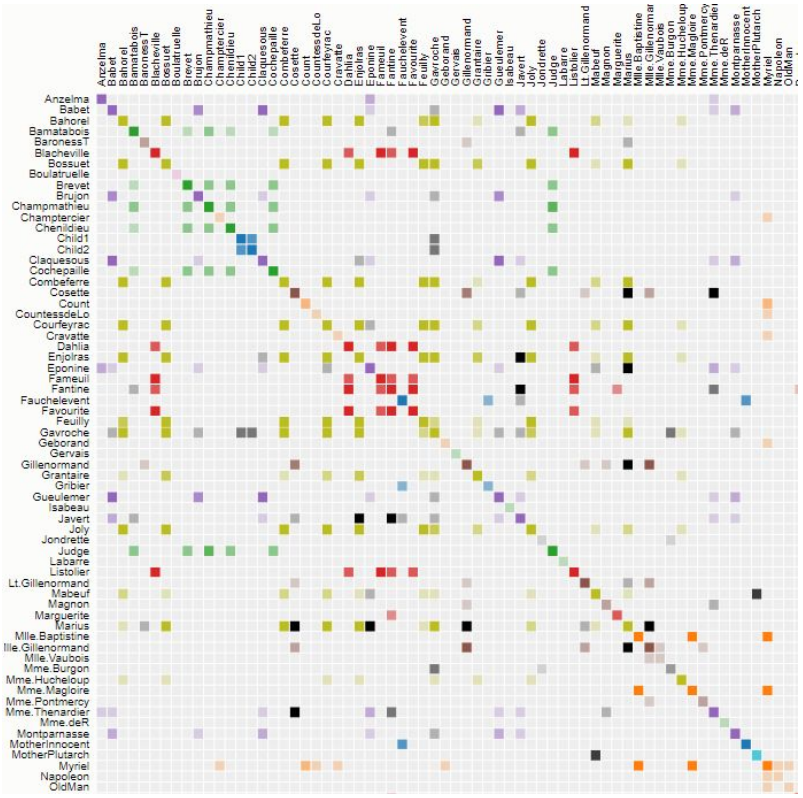
- https://observablehq.com/@d3/brushable-scatterplot-matrix La matrice del grafico a dispersione visualizza le correlazioni a coppie per dati multidimensionali; ogni cella nella matrice è un grafico a dispersione. Questo esempio utilizza i dati di Anderson sui fiori di iris nella penisola di Gaspé.



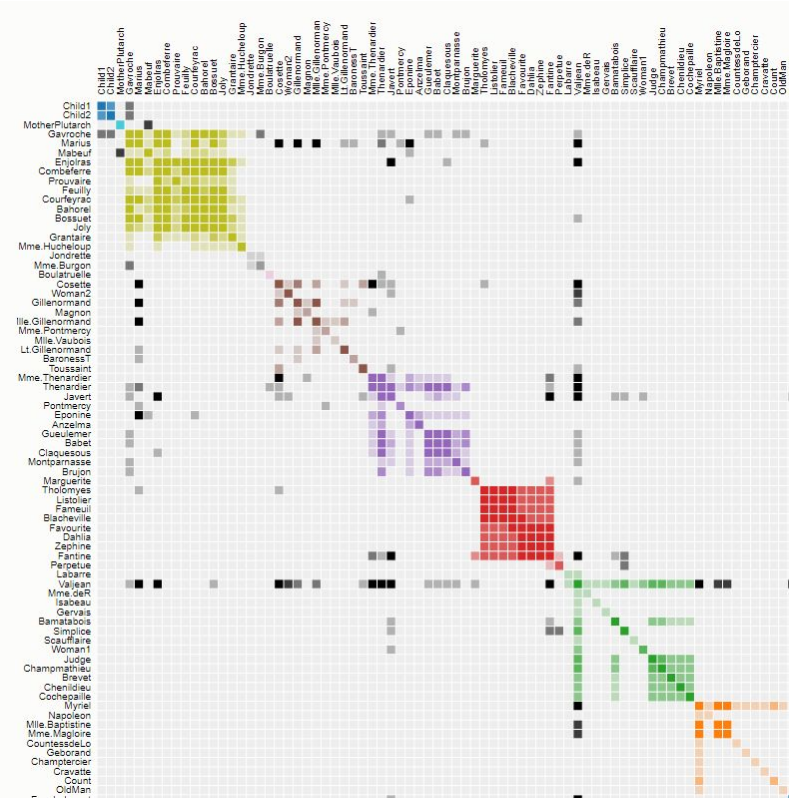
Il grafico è interattivo e selezionando dei valori vengono evidenziati i punti anche negli altri grafici. In questa versione della “Scatter Plot Matrix” nella diagonale si vedono delle linee perché c’è la stessa misura sia sulla x che sulla y, in altre implementazioni nella diagonale (vista la poca utilità della linea) viene disegnato un istogramma che mostra la distribuzione dei valori in quella dimensione.



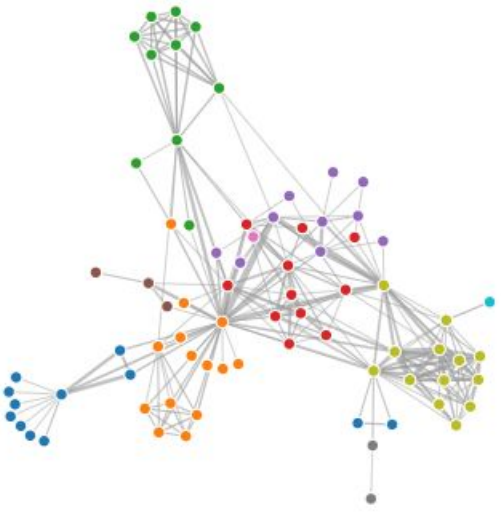
- https://bost.ocks.org/mike/miserables/ Questo diagramma a matrice visualizza le ricorrenze dei personaggi in Les Misérables di Victor Hugo. Ogni cella colorata rappresenta due personaggi che sono apparsi nello stesso capitolo; le celle più scure indicano caratteri che si sono verificati più frequentemente.



come si può notare i punti sono più o meno colorati a seconda del numero di interazioni, cioè della distanza tra un personaggio e l'altro.
Ma questa visualizzazione mantiene i punti dispersi nella matrice, mentre sarebbe molto più utile vedere raggruppati i personaggi che hanno delle forti interazioni tra loro.
Questo si ottiene ordinando i dati secondo appunto il criterio della quantità di interazioni e il risultato si vede nella figura seguente dove si vedono appunto dei quadrati in corrispondenza di raggruppamenti di personaggi.

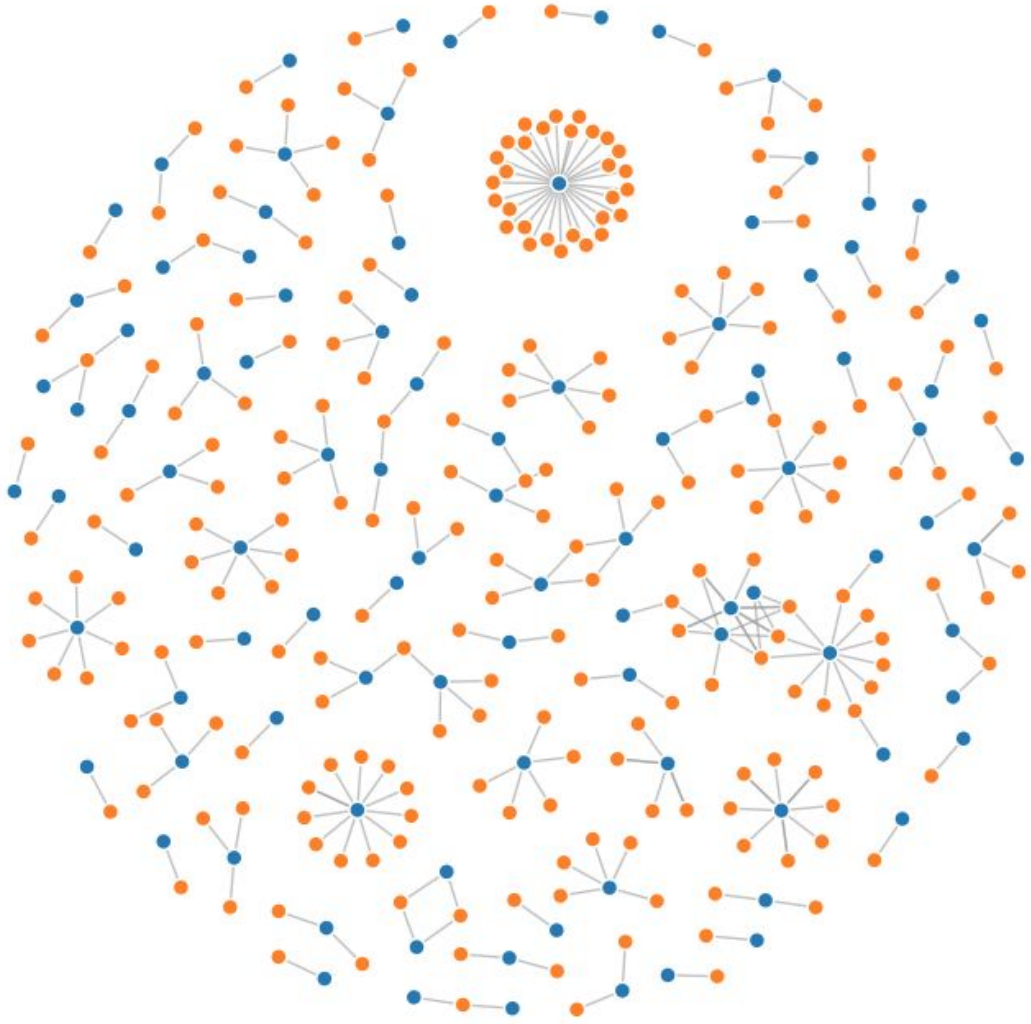


- https://observablehq.com/@d3/force-directed-graph Conosciuto anche come grafico di rete. Mostra come le entità sono interconnesse attraverso l'uso di nodi e linee di collegamento per rappresentare le loro connessioni



In questo caso i punti sono collegati tra loro da un legame che esercita una forza di attrazione, mentre i punti si respingono tra loro.
Il grafico è animato e dopo una prima disposizione casuale i punti si spostano secondo le forze che gli muovono. Dopo un po' viene raggiunta una condizione di minimo che rappresenta un equilibrio stabile e porta ad evidenziare la struttura data dalla geometria multidimensionale alla base della definizione delle forze.

C'è una interessante variazione sul tema: se i punti sono tutti connessi si ha il grafico precedente, invece se ci sono "isole" disconnesse queste tenderanno ad allontanarsi tra loro. Con una ridefinizione delle forze, aggiungendo una forza di attrazione che opera solo a grande distanza, si possono tenere uniti i gruppi disconnessi che altrimenti si allontanerebbero indefinitamente.
Il risultato è quello della figura seguente.



Variazioni ai requisiti

In corso d'opera non sarà possibile variare/modificare i requisiti minimi (obbligatori per accettare il prodotto). Sarà invece possibile variare i requisiti opzionali, in quanto saranno i gruppi vincitori dell'appalto a modificarli / eliminarli / aggiungerli.

Documentazione

Il progetto dovrà essere supportato dalla documentazione minima richiesta per il corso di Ingegneria del software e dovrà essere fornito un manuale per l'utilizzo ed un manuale per chiunque voglia estendere l'applicazione.

Garanzia e Manutenzione

L'azienda Zucchetti SPA è interessata a questo progetto come dimostrazione della fattibilità dell'obiettivo utilizzando le tecnologie web. Costituirà titolo preferenziale nella valutazione delle proposte la pubblicazione del progetto sul sito “github.com” o altri repository pubblici, in conformità con i relativi requisiti di natura open-source, per favorire la continuità del prodotto risultante.

Rinvio

Per tutto quanto non previsto nel presente capitolato, sono applicabili le disposizioni contenute nelle leggi e nei collegati per la gestione degli appalti pubblici.