UNIVERSITÀ DI PISA

Corso di Laurea in Informatica Umanistica

LAUREA TRIENNALE

Visualizzare l'informazione dei social media

Candidato: *Stefano Misuraca*

Relatore: *Maurizio Tesconi*

Correlatore: *Matteo Abrate, Fabio Valsecchi*

Anno Accademico 2015-2016

Sommario

I social media sono ormai una parte integrante della vita di molte persone, vengono infatti usati per comunicare, condividere foto, video, contenuti, trovare informazioni e creano una rete mondiale in cui ogni giorno fluiscono milioni di dati.
La quantità di informazioni prodotta da questi media può essere usata in svariati modi, sia per condurre ricerche statistiche che per approfondimenti di carattere sociale nel campo del Web fino ad arrivare ad elaborazioni più avanzate come l'analisi dei contenuti per comprendere gli eventi che stanno accadendo in una certa zona geografica. Gli utenti diventano quindi veri e propri sensori sociali che forniscono in tempo reale informazioni sulla realtà che li circonda.
Questa tesi si prefigge di illustrare sia la genesi e l'evoluzione dei social media che lo sviluppo delle tecniche per il trattamento di grandi quantità di dati, fino ad arrivare alla descrizione della realizzazione e del funzionamento del Replayer.
L'applicazione realizzata presso il CNR di Pisa consente di analizzare eventi importanti e di attualità, come gli attentati di Parigi o le alluvioni in Italia, attraverso i dati provenienti dai social media e di visualizzare i contenuti in modo dinamico e intuitivo mediante tecniche di visualizzazione dell'informazione.

Indice

Elenco delle figure

Capitolo 1

Social Media e Social Sensing

1.1 Introduzione

1.1.1 Nascita del Web

Nel 1991 il ricercatore Tim Berners-Lee sviluppò presso il CERN di Ginevra[1] il protocollo HTTP (Hyper Text Transfer Protocol), cioè un sistema per la trasmissione di dati attraverso il Web utilizzando un'architettura composta da client e server. Questo protocollo costituisce il nucleo fondamentale del World Wide Web (WWW), un servizio di internet che cambierà per sempre il mondo[12].

Nei primi anni dopo la sua nascita, il Web era usato principalmente dai ricercatori svizzeri per comunicare tra loro e veniva sfruttato come una piattaforma collaborativa per lo sviluppo dei progetti, Berners-Lee affascinato dalle potenzialità di questa tecnologia capì l'importanza di renderla pubblica e accessibile anche ai non addetti ai lavori[5].

Due anni dopo, nel 1992, faceva la sua comparsa Mosaic, il primo browser web a raggiungere una popolarità internazionale. Questa applicazione, seppur ancora rudimentale, permetteva la visualizzazione e la navigazione di risorse situate nel Web, principalmente documenti ipertestuali nella forma di pagine Web scritte mediante il linguaggio di markup HTML[2][30].

Con l'avvento dei primi browser, il Web si diffuse su scala globale e negli anni seguenti il traffico internet crebbe a dismisura così come la quantità di siti web fruibili. L'apice venne però raggiunto negli anni 2000, infatti tra il 2005 e il 2010 il numero degli utenti connessi raddoppiò fino a sorpassare i due miliardi[20].

1.1.2 L'evoluzione del Web e il World Wide Web Consortium

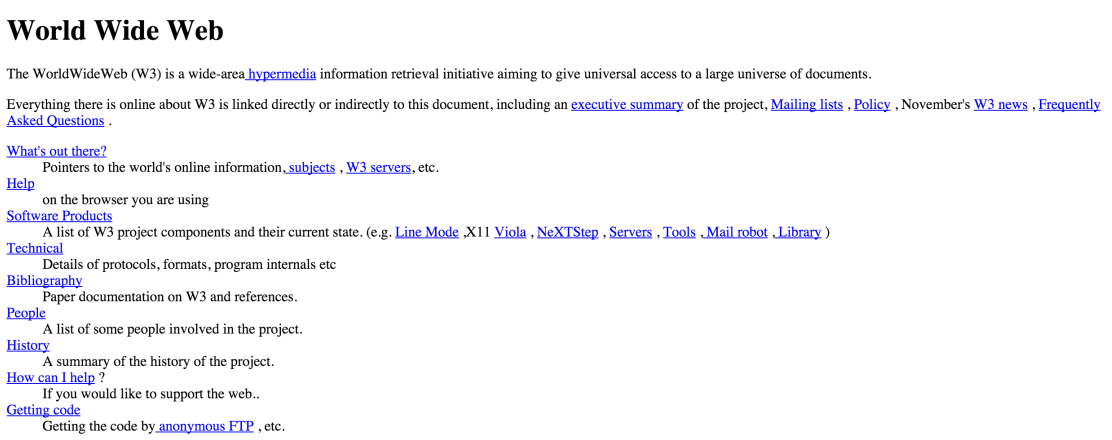
L'aspetto del Web è cambiato molto dalla sua nascita, in origine era infatti utilizzato da pochi utenti ed il suo scopo era principalmente divulgare informazioni tramite pagine web statiche, cioè con un contenuto fisso principalmente testuale, che non permettevano l'interazione con l'utente.

[1] Organizzazione europea per la ricerca nucleare

[2] HyperText Markup Language (HTML) promosso dal World Wide Web Consortium (W3C)

I cambiamenti rispetto al Web odierno sono stati molteplici, l'interattività e la dinamicita sono state possibili grazie all'evoluzione del linguaggio HTML (oggi siamo arrivati alla versione HTML5) e all'introduzione di linguaggi di scripting lato client come JavaScript che permettono, tramite una gestione ad eventi, la creazione di effetti dinamici e interattivi.
Inoltre sono stati migliorati anche i linguaggi lato server, come PHP ASP e JSP, trasformando i web server in veri e propri application server, cioè l'infrastruttura che fornisce il supporto, lo sviluppo e l'esecuzione di applicazioni e di altri componenti[28].

Abbandonato il CERN, Tim Berners-Lee fondò il World Wide Web Consortium (W3C) un'organizzazione internazionale e non governativa con lo scopo di sviluppare tutte le potenzialità del Web e fornire degli standard riconosciuti.
Nel 1998 viene infatti introdotto lo standard XML (eXtensible Markup Language) cioè un metalinguaggio che consente la creazione di nuovi linguaggi di marcatura, con questa innovazione le pagine Web non venivano più considerate solo in base alla loro formattazione HTML ma veniva prestato maggiore interesse anche alla struttura e al significato del contenuto migliorando così il reperimento e la fruizione delle informazioni contenute nel Web[9].



World Wide Web

The WorldWideWeb (W3) is a wide-area hypermedia information retrieval initiative aiming to give universal access to a large universe of documents.

Everything there is online about W3 is linked directly or indirectly to this document, including an executive summary of the project, Mailing lists , Policy , November's W3 news , Frequently Asked Questions .

What's out there?
Pointers to the world's online information, subjects , W3 servers, etc.
Help
on the browser you are using
Software Products
A list of W3 project components and their current state. (e.g. Line Mode ,X11 Viola , NeXTStep , Servers , Tools , Mail robot , Library)
Technical
Details of protocols, formats, program internals etc
Bibliography
Paper documentation on W3 and references.
People
A list of some people involved in the project.
History
A summary of the history of the project.
How can I help ?
If you would like to support the web..
Getting code
Getting the code by anonymous FTP , etc.

Figura 1.1: Primo sito Web della storia creato da Tim Berners-Lee il 6 agosto 1991, ancora visibile all'indirizzo TheProject.html.

1.1.3 Il Web semantico

Il Web è ancora carente di un meccanismo univoco e condiviso per elaborare automaticamente le informazioni situate nelle varie pagine e metterle in relazione tra loro in base al contenuto, all'organizzazione e ad altri criteri di ricerca.
È stato quindi coniato il termine Web Semantico, che sottolinea la trasformazione, ancora in corso, del World Wide Web in un ambiente dove tutti i contenuti sono associati ad informazioni e metadati che ne specificano il contesto semantico, questa espressione

è stata introdotta la prima volta da Tim Berners-Lee in un articolo pubblicato per il Scientific American ed esprime una visione del Web come un insieme di dati processabili dalle macchine.

Tramite il markup semantico e l'utilizzo dei metadati sono possibili ricerche molto evolute basate sulla presenza nel documento di parole chiave e la costruzione di una rete di relazioni logiche (link) tra i nodi presenti all'interno di un contenuto Web.
L'obiettivo principale del Semantic Web è consentire alle macchine di estrarre dal web sia le informazioni che il loro significato, spesso questi dati si trovano in formati eterogenei ma, sfruttando le tecnologie consigliate dal W3C (RDF, SPARQL) è possibile elaborarli per estrapolare i contenuti[31].

Interpretando i contenuti secondo le linee guida del Web Semantico e creando una metodologia indipendente dallo specifico ambiente operativo è possibile raffinare l'interrogazione e l'interpretazione da parte dei motori di ricerca e di programmi per l'elaborazione automatica e gestire quindi al meglio la grande quantità di informazioni contenuta nel Web e la sua fruizione[6].

1.1.4 Il World Wide Web oggi: Web 2.0

Oggi secondo le stime dell'International Telecommunication Union[3] gli utenti che utlizzano regolarmente il Web sono circa 3 miliardi, ovvero più del 35% della popolazione mondiale. Un aspetto interessante è l'enormità di dati che vengono creati e conservati, le stime infatti indicano che ogni secondo vengono generati 32GB di dati in tutto il mondo e molti di questi provengono dai social network[27].

Si parla quindi di Web 2.0, cioè uno stato dell'evoluzione del Web che comprende l'avvento a la diffusione mondiale di tutte le applicazioni online che permettono un alto livello di interazione tra il sito e l'utente, lasciandosi alle spalle definitivamente la staticità del Web 1.0.
L'interattività e la dimanicità del Web hanno preso forma in blog, wiki, social network, e-commerce e tutte le piattaforme per la condivisione di media di vario genere che sfruttiamo ogni giorno[26].

1.2 Social Media

1.2.1 Introduzione

Ogni giorno milioni di persone utilizzano i social media per comunicare tra di loro, ottenere informazioni e compiere una vasta gamma di attività, ma dare una definizione precisa di social media può risultare difficile, ancora più arduo è distinguere quali programmi o applicazioni rientrano tra i social media e quali no.
Andreas Kaplan e Michael Haenlein hanno condotto numerose ricerche sui media moderni e il marketing virtuale fino ad arrivare a definire i social media come: "Un gruppo

[3]http://www.itu.int/ Agenzia delle Nazioni Unite che si occupa delle problematiche riguardanti l'informazione e le tecnologie di comunicazione

di applicazioni internet che si basano su presupposti ideologici e tecnologici del Web 2.0 e che consentono la creazione e lo scambio di contenuti generati dagli utenti".
I due studiosi utilizzano la dicitura Web 2.0 per descrivere una nuova modalità di utilizzo del Web, cioè una piattaforma in cui contenuti e applicazioni non vengono più creati e pubblicati da individui singoli, ma al contrario sono continuamente modificati da tutti gli utenti in modo collaborativo e partecipativo.

Oltre a questa definizione generale è opprtuno identificare una classificazione che distingua i social media da altri prodotti del Web, molte persone sarebbero probabilmente d'accordo nell'includere YouTube, Facebook e Second Life[4] nel gruppo social, ma non esiste un sistema specifico con cui i diversi social media possono essere categorizzati e che includa non solo quelli attuali ma anche quelli che potrebbero essere generati in futuro.
I due studiosi si sono quindi basati su un'insieme di teorie nel campo della ricerca sui media, in particolare quella sulla *presenza sociale* e la *ricchezza dei media* e su altri processi sociali coinvolti poichè questi sono gli elementi chiave per comprendere la natura e l'evoluzione dei social media[17].

1.2.2 Teoria della presenza sociale e ricchezza dei media

La teoria della presenza sociale è stata sviluppata da John Short, Ederyn Williams e Bruce Christie e classifica diversi media di comunicazione rispetto alla continuità della presenza sociale, dove il grado di presenza è equiparato al grado di consapevolezza della persona durante l'interazione comunicativa[15].

Secondo questa teoria la comunicazione è efficace se il media utilizzato ha un grado adeguato di presenza sociale cioè garantisce un buon livello di coinvolgimento interpersonale. La presenza sociale viene incrementata dal contatto acustico e visivo tra le persone che stanno comunicando e viene incentivata con l'intimità e l'immediatezza del media utilizzato. Ci si aspetta infatti che la presenza sociale sia minore nei media asincroni come l'email rispetto a quelli sincroni e immediati come le chat oppure in quelli in cui la comunicazione è mediata da un apparecchio elettronico come il telefono rispetto alla conversazione faccia a faccia.
Maggiore è la presenza sociale e più grande sarà l'influenza che le parti in comunicazione hanno sul reciproco comportamento[17].

Correlata all'idea della presenza sociale c'è il concetto di ricchezza del media utilizzato. Questa teoria è stata fomulata la prima volta da Daft e Lengel nel 1986 e si basa sull'assunzione che l'obbiettivo di ogni comunicazione è la risoluzione di ambiguità e la riduzione dell'incertezza.
I media differiscono quindi nel grado di ricchezza che possiedono, cioè nella quantità di informazioni che permettono di scambiarsi all'interno di un dato intervallo di tempo e di conseguenza alcuni media sono più efficaci di altri nel risolvere ambiguità e incertezze.

[4]Second Life è un mondo virtuale online sviluppato da Linden Lab (San Francisco) e pubblicato nel 2003

Applicando queste teorie al concetto dei social media possiamo assumere che una prima classificazione può essere realizzata sulla base della richezza e della presenza sociale che quel media dispone[10].

Figura 1.2: Classificazione dei social media più popolari in base alla loro tipologia.

1.2.3 Diffusione dei social network

All'interno dei social media vengono raggruppati i social network, cioè strumenti che consentono lo sviluppo di una rete di persone interconnesse tra loro da diversi legami sociali, si parla quindi di una vera e propria rete sociale virtuale che si basa su legami di conoscenza, parentela, interessi comuni o vincoli lavorativi[11].

Questi tipi di media nascono poco dopo il 2003, sebbene già nel 1997 si potessero trovare alcuni prototipi, tra i primi ricordiamo Orkut, LinkedIn ed MSN, quest'ultimo di particolare successo in Italia. Sin dai primi anni questi servizi hanno registrato un notevole traffico ma fu solo nel 2006 che si ebbe un incremento enorme di iscritti con l'introduzione del social network più conosciuto al mondo: Facebook. La piattaforma conta oggi circa 1.40 miliardi di iscritti e ogni minuto vengono caricate 136 mila foto, 300 mila stati vengono aggiornati e vengono scritti migliaia di commenti[21].

Un aspetto fondamentale dell'analisi dei social network è infatti la viralità, intesa non solo come la diffusione del media stesso, ma come la probabilità che l'utente condivida

un contenuto generato da un altro utente all'interno della sua rete sociale. Facebook è stato tra i primi ad introdurre strumenti specifici per aumentare la viralità come le funzioni di condivisione, la possibilità di mettere 'Mi Piace' ai contenuti e di inserire commenti. Infatti più alta è l'interazione tra gli utenti e la condivisione dei contenuti e maggiore è il successo del social network, fino ad arrivare a dare vita a veri e propri fenomeni sociali[1].

1.2.4 Distinzione da altri media

Parlando di social media è naturale associarli ai media tradizionali, come televisione, cinema e radio. Infatti, per comprendere al meglio che cosa sono e in cosa consiste questa nuova categoria di media, è importante mettere in luce le differenze e i punti in comune con i mezzi di comunicazione di massa.

Una delle principali differenze è sicuramente l'accessibilità, i social media hanno infatti un costo per l'utente pressoché nullo o relativamente basso. Questo permette a chiunque di accedervi e di usarli per pubblicare qualsiasi tipo di dato o informazioni, a differenza dei media tradizionali che invece necessitano di abbonamenti o sottoscrizioni di vario tipo per poter usufruire delle informazioni e dei contenuti proposti.
Un'altra caratteristica che differenzia le due categorie è l'usabilità, intesa come la facilità di utilizzo di queste tecnologie. I media industriali, come la televisione, richiedono tipicamente delle abilità specifiche e una formazione adeguata per poter creare dei contenuti, al contrario nella maggior parte dei social media sono sufficienti abilità modeste e attrezzature comuni per diventare creatori di informazioni e contenuti. Basti pensare agli utenti che generano video su YouTube in contrapposizione a chi produce film e telefilm per un canale televisivo o per il grande schermo.
Una caratteristica che invece accomuna questi media è il numero degli utenti finali, cioè coloro che possono usufruire del contenuto proposto. Che si tratti di un tweet o di un post su Facebook oppure di una trasmissione televisiva, tutti questi contenuti hanno una portata mondiale.

Le ultime differenze tra i media tradizionali e i social media riguardano l'immediatezza e la permanenza dei contenuti, con il primo concetto si intende il tempo che intercorre tra le comunicazioni, infatti in un media tradizione possono passare giorni, settimane e in alcuni casi mesi, mentre nei social media la risposta può essere istantanea e seguire i fatti in tempo reale. Per quanto riguarda la permanenza invece, nei media industriali un contenuto, dopo che è stato creato, non può essere alterato, come l'articolo di un giornale dopo la sua distribuzione o un programma dopo la sua messa in onda, mentre alla base dei social media risiede la cooperazione tra gli utenti della creazione e nell'ampliamento dei contenuti proposti[19].

1.3 Social Sensing

1.3.1 Introduzione

La proliferazione dei social network ha incrementato la consapevolezza riguardo il potere dei dati generati da questi media, infatti molte applicazione sono state incorporate con elementi sociali e interazioni con i social media. Questi network contengono solitamente una gamma eterogenea di dati che possono essere usati in svariati modi.

Per migliorare la potenza di tali applicazioni si possono incorporare all'interno delle stesse dei sensori che raccolgono continuamente grandi quantità di dati che possono essere poi usati per operazioni di previsione e monitoraggio. Questo ha portato alla nascita di numerosi sistemi di social sensing come Biketastic, CarTel e SoS Project[5] che tengono traccia degli spostamenti e delle attività degli utenti, quest'ultimo sviluppato presso il CNR di Pisa. Nel caso di Biketastic le persone possono indicare nuovi percorsi da svolgere in bicicletta, usufruire di una mappa interattiva e condividere con gli altri i percorsi che hanno scoperto. Tramite SoS Project e l'analisi dei dati provenienti dai social media è stato possibile monitorare e visualizzare un evento catastrofico della storia italiana, il terremoto in Emilia del 2012. Il progetto si occupa sopratutto di studiare i comportamenti sociali in risposta al verificarsi di calamità naturali, in modo particolare di terremoti[3].

I sensori sociali si possono sfruttare non solo per reperire informazioni geografiche da parte degli utenti, ma anche per migliorare la quantità e qualità di dati su un argomento, infatti permettendo agli utenti dei social network da un lato di pubblicare le loro informazioni e dall'altro di poter accedere a quelle pubblicate dagli altri viene incrementata la consapevolezza in tempo reale di quello che accade intorno a loro e fornisce delle informazioni riguardo il comportamento globale nei confronti di un fenomeno.

Sfruttando i dati prodotti dalle persone, che diventano quindi i sensori sociali in prima persona, si possono condurre ricerche per comprendere i meccanismi di aggregazione di alcune comunità o gruppi in relazione all'ambiente esterno, ad esempio identificare le condizioni del traffico di una città oppure misurare la tendenza all'obesità o ancora capire la percezione che hanno gli utenti riguardo l'inquinamento.

Tutti questi servizi, da quelli che includono il tracciamento della posizione, la misurazione dell'attività fisica, il geo-tagging o l'individuazione di punti di interesse da parte degli utenti, hanno causato un cambiamento nei paradigmi computazionali che viene definito come crowd-sourcing e indica cioè il coinvolgimento attivo della popolazione nel meccanismo di collezione ed elaborazione dei dati[2].

1.3.2 Problematiche del social sensing

Con l'avvento del social sensing i ricercatori si sono trovati ad affrontare una serie di problematiche, la prima tra queste riguardante la privacy, infatti i dati raccolti dai social media contengono spesso informazioni personali ed è quindi importante tenere sotto controllo la sicurezza degli utenti e delle loro informazioni. Negli ultimi anni sono

[5] http://socialsensing.it/ SoS Project

state quindi realizzate delle tecniche per collezionare e usare i dati provenienti dai sensori sociali senza intaccare la privacy degli utenti. PoolView è una tra le tecniche sviluppate più recentemente e si basa su una serie di operazioni matematiche e statistiche sui dati al fine di garantirne l'anonimato.
In particolare viene usata la tecnica della Data Perturbation (perturbazione dei dati) che consiste nell'applicare tecniche di distribuzione di probabilità e poi effettuare operazioni di sostituzione dei dati, l'altro approccio comune è quello della distorsione che applica una perturbazione nei dati aggiungendo o moltiplicando il rumore presente oppure creando dei processi randomici[13].

Altri fattori critici da considerare in questo campo sono l'utilizzo delle risorse, infatti i sensori sociali, specialmente se implementati all'interno di dispositivi mobili o indossabili (smart-watch, fitbit), funzionano mediante l'uso di batterie che potrebbero incidere in maniera negativa sulla durata della carica del dispositivo. Bisogna quindi trovare il giusto bilanciamento tra lo scopo dell'applicazione e il consumo di risorse.
Inoltre, sempre riguardo le risorse impiegate, si deve considerare la mole dei dati raccolti, basti pensare alle applicazioni che tengono traccia della localizzazione e che ogni secondo memorizzano grandi quantità di informazioni.

Considerando che i dati raccolti con il social sensing provengono da sensori oppure sono inseriti direttamente dagli utenti, è bene considerare l'attendibilità delle informazioni raccolte, sia per calcolare gli errori tecnici dovuti all'accuratezza dei sensori stessi sia per gli errori umani. Inoltre gli schemi che preservano la privacy riducono ancora di più la fedeltà dei dati in modo considerevole perché non si può avere una traccia precisa della tipologia di utente che li ha inseriti[32].

Capitolo 2

Information Visualization: visualizzare grandi quantità di dati

2.1 Big Data

2.1.1 Introduzione

Il termine Big Data viene utilizzato quando ci si riferisce a dataset (collezione di dati) talmente estesi da richiedere sistemi specifici per estrapolare, gestire e processare le informazioni contenute. Infatti utilizzando le tecniche convenzionali non si avrebbero dei risultati apprezzabili sia in termini di tempo di elaborazione che di consumo di risorse hardware.

Big Data è un termine generico e non indica una quantità specifica di dati, poiché le informazioni ed i sistemi che le elaborano si modificano e crescono nel tempo è difficile definire un numero che stabilisca l'appartenenza alla categoria dei Big Data. In generale i dati trattati rientrano nell'ordine di Zettabyte, ovvero miliardi di Terabyte, proprio per questo motivo è richiesta una potenza di calcolo enorme, spesso distribuita su decine, centinaia o addirittura migliaia di server.

I big data si possono accomunare tramite alcune caratteristiche:

- Volume: la dimensione effettiva del dataset, questa misura incide nel calssificare i big data come tali.
- Velocità: intesa come velocità di generazione e di elaborazione dei dati.
- Varietà: le varie tipologie di dati che possono provenire da fonti diverse.
- Variabilità: si riferisce alla possibile inconsistenza dei dati che può generare problemi nell'elaborazione.

- Complessità: la qualità dei dati acquisiti, può variare notevolmente e influire su una buona analisi delle informazioni.

Queste grandi quantità di dati vengono memorizzate in enormi database ed utilizzano un modello di struttura che li differenzia dai tradizionali database relazioni (RDBMS Relational DataBase Management System) come MySQL a favore di modelli non relazionali come NoSQL che vengono implementati secondo diverse metodologie come ad esempio: database orientati al documento o alla struttura chiave-valore (MongoDB e Cassandra).

Tenendo in considerazione che oggi nel mondo ci sono attivi quasi 4 miliardi di smartphone, più del 35% della popolazione ha accesso ad internet e considerando inoltre che la maggior parte degli utenti del Web utilizza i social network abitualmente, non è difficile intuire che la quantità di dati che vengono generati da piattaforme come Facebook o Twitter è colossale[21].

2.1.2 Reperire e utilizzare i big data

Le applicazioni basate sull'analisi di big data stanno diventando sempre più popolari, soprattutto perché data la grande quantità di informazioni, i possibili utilizzi sono svariati e applicabili a tutte le scienze. L'analisi di questi dataset permette infatti di individuare delle tendenze nell'area finanziaria, fare delle stime sulla diffusione delle malattie e sulla loro prevenzione, avere informazioni sulla criminalità e altri fattori sociali.

Il numero di dataset disponibili sta rapidamente aumentando anche grazie a tutti i dispositivi elettronici che producono una grande quantità di informazioni in maniera rapida e senza costi, come ad esempio i log dei software, tracciati di veicoli e dispositivi wireless. Per comprendere al meglio la portata di questi dati, si consideri che ogni giorno vengono creati 2.5 exabyte di nuovi dati e ogni 40 mesi raddoppiamo la nostra capacità tecnologica di memorizzare informazioni.

Nel campo delle telecomunicazioni si parla di Internet of Things (internet delle cose) per indicare la possibilità di estendere l'accesso ad internet anche agli oggetti. Così facendo questi strumenti elettronici avrebbero la possibilità di comunicare dati o ricevere informazioni da altri dispositivi. Ad esempio la sveglia potrebbe suonare in anticipo in caso di condizioni di traffico elevato, in questo progresso tecnologico i big data sono un punto fondamentale, infatti servono a gestire la connettività di tutti questi dispositivi, le informazioni che producono ed elaborano e possono calibrare al meglio l'ottimizzazione di queste tecnologie.

L'attenzione verso i big data ha raggiunto un livello mondiale, molti governi stanno investendo denaro nella ricerca su questi enormi dataset, nel 2012 l'amministrazione Obama ha istituito un'iniziativa per esplorare le possibilità di utilizzo dei big data per fronteggiare alcuni problemi governativi e i big data hanno giocato un ruolo fondamentale nella previsione dei risultati delle elezioni presidenziali e nello stilare le statistiche di gradimento da parte delle popolazione[25].

Nel campo tecnologico l'utilizzo dei big data è ormai consolidato per citare qualche esempio eBay, la famosa piattaforma di e-commerce, sfrutta 7.5 petabytes di dati per

raccogliere e analizzare i commenti degli utenti, le ricerche e tutto ciò che ha a che fare con gli acquisti. Amazon invece nel 2005 ha detenuto il record per i tre più grandi database in funzione in tutto il mondo[18], anche i social network sfruttano abbondamente i big data, Facebook lavora infatti con 50 miliardi di foto caricate dagli utenti e Google raggiunge i 100 miliardi di ricerche al mese.

2.1.3 Machine Learning

Il Machine Learning (apprendimento automatico) è un ambito di ricerca dell'informatica che si è sviluppato partendo dagli studi sull'intelligenza artificiale riguardanti il riconoscimento di pattern e l'apprendimento computazionale.
Arthur Samuel, pionere in questo campo, ha definito il Machine Learning come un settore di studi che fornisce ai computer l'abilità di imparare qualcosa, senza essere esplicitamente programmati per farlo[29].

L'apprendimento può avvenire elaborando le caratteristiche principali provenienti da esempi, strutture dati o sensori, le macchine devono quindi analizzarle e valutare le relazioni tra le variabili osservate.
Ai computer viene fornito un insieme di esempi già elaborati manualmente, chiamati dati di allenamento e vengono poi scritti dei programmi che partendo dall'analisi di questi insiemi devono riconoscere all'interno di un nuovo set di dati l'esempio più simile a quello di partenza.

Uno degli obiettivi principali della ricerca sull'apprendimento automatico è quello di insegnare alle macchine a riconoscere automaticamente modelli complessi e prendere decisioni intelligenti basate sui dati. La difficoltà risiede nel fatto che l'insieme di tutti i possibili comportamenti è troppo grande per essere coperto da insiemi già osservati. Da qui è necessario l'utilizzo di tecniche per generalizzare i vari casi, in modo da essere in grado di produrre un comportamento utile per le nuove situazioni.

Esistono due tipi di algoritmi di Machine Learning:

- Algoritmi di apprendimento supervisionati: Nell'apprendimento supervisionato i dati di allenamento sono composti da una coppia di esempi formata da un oggetto di input e il valore di output desiderato. Questo tipo di algoritmo deve essere in grado di predire il corretto valore di output per ogni input fornito e quindi si presta meglio per risolvere i problemi di classificazione.

- Algoritmi di apprendimento non supervisionati: Nell'apprendimento non supervisionato non è necessario utilizzare dei dati di allenamento, questi algoritmi sono particolarmente efficienti nei problemi di ranking, ovvero ordinare in base ad qualche proprieta dei dati, come il valore o la grandezza e il clustering, cioè raggruppare i dati secondo la loro similarità.

Oggi questi algoritmi trovano un vasto uso nel campo della sintesi vocale oppure nei lavori di traduzione automatica dei testi, sono molto utilizzati anche nel campo dei social media e dei big data poiché questi riescono a rivolere problemi come raggruppamenti o classificazioni in maniera automatica lavorando con enormi quantità di dati. Bisogna

però considerare che non sempre questi algoritmi riescono a portare a termine i loro processi senza produrre errori, bisogna infatti valutare bene se l'eventuale correzione manuale degli errori è giustificata dalla percentuale di successo della macchina[14].

2.2 Information Visualization

L'Information Visualization, spesso abbreviata in InfoVis, è uno studio che ha come scopo la rappresentazione di dati astratti per facilitare la comprensione da parte dell'utente e migliorare la sua capacità cognitiva. I dati possono essere di qualunque tipo, per esempio numerici, non numerici o geografici. Visualizzare i dati è fondamentale per una buona comprensione, le industrie e le aziende basano i loro studi di mercato su grafici che rappresentano le vendite o le previsioni così da percepire con facilità l'andamento del mercato.

L'InfoVis trova le sue origini dalla Computer Grafica, che è sempre stata utilizzata per scopi scientifici e di ricerca, essa infatti permetteva, tramite l'uso di processori dedicati, di elaborare immagini virtuali allo scopo di rappresentare grafici o visualizzazioni complesse. La scarsa potenza delle macchine ha limitato molto lo studio sulle visualizzazioni, fu solo dopo gli anni 80 che, con l'avvento di nuove tecnologie e il maggiore interesse da parte della comunità scientifica, vennero organizzate numerose conferenze e seminari riguardo le tecniche di visualizzazione, dando origine così al processo di sviluppo e innovazione che ci ha condotto fino ad oggi.

2.2.1 Cosa, come e perché visualizzare

Quando si devono rappresentare dei dati spesso è difficile identificare le tecniche di visualizzazione migliori e in passato sono stati creati molti strumenti semi-automatici o metodologie di presentazione delle informazioni allo scopo di aiutare gli sviluppatori. Poiché esistono svariate combinazioni di dati, finalità di utilizzo delle rappresentazioni e categorie di utenti destinatari rimane ancora una sfida identificare la visualizzazione più corretta.
Esistono molti modi di affrontare un problema di visualizzazione, in primo luogo bisogna interrogarsi su tre domande chiave:

- Quali sono i dati che l'utente sta analizzando?
- Perché l'utente necessita di uno strumento di visualizzazione?
- Come sono costruiti la codifica visiva e l'interazione con i simboli presentati per quanto riguarda le scelte di design?

Rispondere a queste domande può essere d'aiuto per capire la natura del problema, trovare una soluzione efficace e scegliere il giusto tipo di visualizzazione da applicare.
Quando le collezione di dati sono molto estese esistono dei fattori che limitano la comprensione dei dati, uno è la percezione dell'utente che, non riuscendo a visualizzare o a comprendere informazioni molto complesse o articolate non può focalizzare l'attenzione sulla totalità dei dati o apprezzarne i particolari.

Un'altra limitazione è il display, cioè lo strumento sul quale si visualizzano i dati che, a causa delle sue dimensioni o proprietà (la risoluzione dello schermo per esempio), non riesce a presentare correttamente la complessità o la dimensione di dati in esame. La soluzione è dunque creare visualizzazioni interattive dove l'utente può, tramite delle azioni, modificare quello che sta osservando per poter comprendere i dati in maniera generale ma anche le relazioni che intercorrono tra essi[24].

Prima della diffusione di moderne tecnologie legate alla computer grafica, le visualizzazioni erano limitate ad immagine statiche stampate su fogli, ma con le visualizzazioni generate dai computer oggi possiamo renderli interattivi aumentando di molto le potenzialità dell'InfoVis.

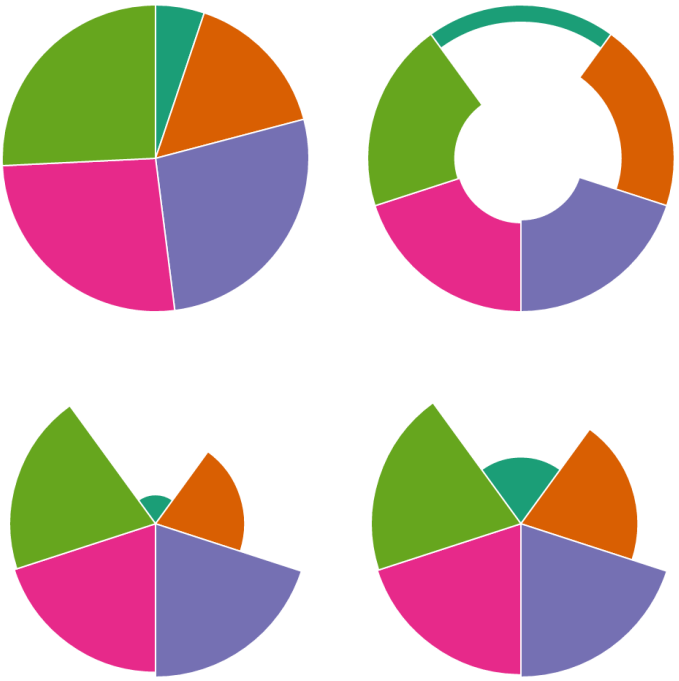


Figura 2.1: Esempio di visualizzazione delle stesse quantità di dati tramite grafici diversi.

2.2.2 Creare visualizzazioni

Partendo da un set di dati, il processo di creazione di una visualizzazione può essere molto lungo e complesso, solitamente si articola in quattro fasi che possono essere ripetute più volte prima di arrivare al risultato finale.
Il primo passo è raccogliere e memorizzare i dati, questo passaggio include il controllo dei dati stessi per rintracciare eventuali errori e controllarne la coerenza, successivamente vi è una fase di pre-elaborazione dove le informazioni vengono trasformate in qualcosa di comprensibile per l'essere umano e quindi i numeri e i dati vengono ordinati o raggruppati per formare delle serie da presentare.

Successivamente entrano in gioco hardware, tecniche di presentazione e algoritmi di grafica per produrre una visualizzazione tramite un supporto informatico, in questa fase si sceglie come visualizzare le informazioni, quanti livelli di analisi fornire e il livello di interattività dell'applicazione.
Alla fine di questo processo tutta la visualizzazione è affidata alla percezione dell'utente che sfruttando il suo sistema cognitivo comprende i dati presentati e amplia la sua conoscenza, è quindi importante analizzare la tipologia di utenti a cui è destinata la visualizzazione per indentificare le tecniche di creazione migliori. [33]

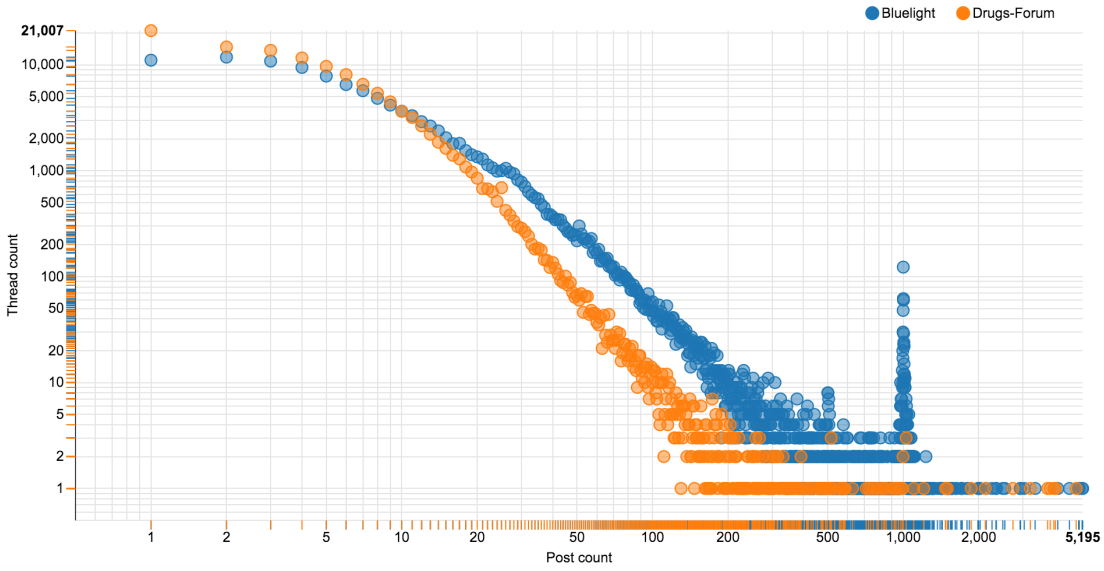


Figura 2.2: Visualizzazione complessa che rappresenta il numero di post all'interno di due forum che trattano lo stesso argomento.

2.2.3 Strumenti per le visualizzazioni

Data la grande diversità e complessità dei dati che possono essere visualizzati esistono strumenti che si adattano meglio a certe tipologie di informazioni. Infatti presentare numeri, relazioni oppure testi scritti pone delle sfide diverse, però la maggior parte dei dati si presenta come un insieme di variabili oppure altre categorie di informazioni possono essere ricondotte a questa forma per poi essere elaborate tramite strumenti di codifica delle informazioni.

Le informazioni geografiche, ad esempio, possono essere rappresentate utilizzando delle mappe che codificano le informazioni che si vogliono visualizzare tramite l'uso di colori o di simboli, mentre la maggioranza dei dati numeri si possono presentare mediante i grafici, ne esistono diverse tipologie principalmente si dividono in:

- Grafici composti da aree le cui ampiezze sono proporzionali alla frequenza di un insieme di dati, come i diagrammi circolari (grafici a torta) o gli istogrammi.
- Grafici che permettono di posizionare una serie di variabili su uno spazio cartesiano, come i grafici a dispersione.
- Grafici che rappresentano strutture come alberi o grafi e utilizzano un insieme di elementi (nodi) che vengono collegati tra di loro mediante delle linee (archi) per presentare il rapporto gerarchico tra i dati.

Ogni giorno vengono sperimentate nuove tecniche di visualizzazione, incluse quelle che sfruttano il 3D o la realtà aumentata e quindi in futuro saremo in grado di produrre visualizzazioni sempre più complesse e interattive[7].

Capitolo 3

Programmazione e grafici dinamici

È molto importante progettare con cura una visualizzazione, in modo particolare se si tratta di un contenuto dinamico ed interattivo, queste visualizzazioni sono oggi le più diffuse poiché permettono di esprimere una complessità di dati maggiore.
Per visualizzare in maniera efficiente ed efficace la maggior parte dei dati si può ricorrere ad uno strumento di visualizzazione molto utilizzato: il grafico, ne esistono infatti svariate tipologie che permettono di visualizzare la gran parte dei dataset, siano essi di carattere numerico, geografico oppure testuale.

I computer sono molto d'aiuto nella progettazione e nella costruzione di un grafico, esistono infatti molti programmi che facilitano il lavoro dello sviluppatore mettendo a disposizione template[1] o layout[2] predefiniti con dei grafici di partenza che successivamente vengono modificati e adattati per rispecchiare i dati in esame.
Purtroppo questi programmi non sono sempre sufficienti o non sono adatti a risolvere tutti i problemi di visualizzazione, per questo motivo, parallelamente a programmi con interfaccia grafica, si sono sviluppati linguaggi di programmazione, librerie e framework per permettere ai creatori di visualizzazioni di disegnare un grafico tramite la programmazione[23].

3.1 Programmazione e WYSIWYG a confronto

WYSIWYG è un acronimo che si estende in: What You See Is What You Get, ovvero "quello che vedi è quello che ottieni". Questo paradigma si basa sul principio che il contenuto che viene inserito tramite l'interfaccia del programma (immagini e testo) avrà lo stesso aspetto di quello stampato in output dal software. Per esempio, il famoso programma di videoscrittura Word è un word processor[3] che si basa sul paradigma WYSIWYG, infatti l'utente si aspetta che tutto il lavoro svolto con l'ausilio del programma

[1] Template: Termine inglese che può essere tradotto come: modello, schema, struttura base
[2] Layout: impaginazione
[3] Programma di elaborazione e formattazione di testi

verrà presentato esattamente così come scritto in origine[16].
Al contrario il paradigma di programmazione si basa sulla conoscenza di linguaggi di programmazione o di markup come strumento di lavoro, questi linguaggi hanno una propria sintassi e una serie di regole che li governano. Il programmatore utilizza quindi un linguaggio per creare dei programmi che, al momento della loro esecuzione, restituiscono l'output desiderato.

Il vantaggio principale dei programmi WYSIWYG è che l'utente ha un feedback immediato del risultato del lavoro, questo gli permette di modificare in tempo reale l'aspetto finale del progetto, non c'è alcuna attesa tra la modifica del documento e il risultato della modifica stessa.
Un altro vantaggio è che questi programmi non necessitano di nessun tipo di conoscenza specifica come accade invece per un linguaggio di programmazione o di markup, questo permette anche a chi non ha particolari conoscenze informatiche di poter svolgere analisi e creare visualizzazioni[1].

D'altra parte i programmi WYSIWYG soffrono di alcuni difetti, per esempio non danno allo sviluppatore la totale libertà di decisione o non offrono un controllo completo e dettagliato sul progetto in corso. Inoltre se il programma genera un risultato sbagliato, per esempio a causa di un bug[4] nel software, non è possibile risolvere il problema direttamente poiché il codice sorgente solitamente non è accessibile e modificabile.

I linguaggi di programmazione invece sono degli strumenti molto flessibili, il programmatore che li utilizza ha il controllo totale su tutti gli aspetti del programma che sta scrivendo. Il problema dei bug è attenuato poiché lo sviluppatore ha il pieno controllo del codice ed ha quindi la possibilità di correggere gli errori.
Il linguaggio di programmazione però necessita di uno studio approfondito per essere utilizzato ed è difficile da sfruttare appieno per un utente. Inoltre questi linguaggi non danno un feedback diretto allo sviluppatore che deve necessariamente ordinare al calcolatore di compilare il codice che ha scritto ogni volta che questo viene modificato.

Il vantaggio più grande per il quale si dovrebbe preferire l'approccio della programmazione è sicuramente per la possibilità di creare programmi specifici per i dati e le visualizzazioni in esame, così facendo l'applicazione sarà ottimizzata e le componenti interattive e dinamiche potranno essere studiate appositamente per quel set di informazioni[22].

3.2 Data Drive Document

3.2.1 Linguaggi di programmazione per il Web

Le pagine Web che ogni giorno visitiamo sono scritte tramite dei linguaggi di programmazione e di markup. Le tecnologie più comuni sono:

- HTML: linguaggio di markup che definisce la struttura della pagina.
- CSS: linguaggio che definisce lo stile di una pagina (colori, posizionamenti e dimensioni).

[4]Il termine bug o baco, identifica un errore nella scrittura di un programma software

- JavaScript: linguaggio di programmazione lato client che rende interattive le pagine tramite la gestione di eventi.
- PHP: linguaggio di programmazione lato server che permette di generare pagine web dinamicamente.

Come fa un linguaggio di programmazione a generare una visualizzazione composta da elementi grafici? La risposta è l'utilizzo di SVG, Scalable Vector Graphics (Grafica Vettoriale Scalabile), questo standard indica una tecnologia in grado di visualizzare oggetti di grafica vettoriale e di gestire immagini scalabili dimensionalmente. Più specificamente si tratta di un linguaggio derivato dall'XML che si pone l'obiettivo di descrivere figure bidimensionali statiche o animate.

Utilizzando SVG unitamente a JavaScript si è in grado di disegnare figure sullo schermo e utilizzarle poi per costruire una visualizzazione o un grafico che rappresenti i dati che si stanno analizzando.
Creare un grafico utilizzando JavaScript e SVG è comunque impegnativo, bisogna sempre lavorare con migliaia di dati e tenere conto di molti fattori, per esempio le posizioni o la forma degli elementi grafici e spesso la situazione può risultare complessa.

Per questo motivo nasce la libreria D3.js (Data Driven Documents) sviluppata da Mike Bostock nel 2011, D3 è una libreria JavaScript che permette di creare visualizzazioni dinamiche ed interattive partendo da dati organizzati, visibili attraverso un comune browser. Per fare ciò la libreria si serve largamente degli standard Web: SVG, HTML5, e CSS. Diversamente da molti altri strumenti, D3 permette un ottimo controllo sulla resa visiva e sul risultato finale.

3.2.2 Esperimenti con i grafici in D3

La libreria, incorporata in una pagina web HTML, utilizza funzioni JavaScript per selezionare elementi del DOM (Document Object Model), creare elementi SVG, assegnargli uno stile grafico, oppure inserire transizioni, effetti di movimento o tooltip.
Questi oggetti posso essere largamente personalizzati utilizzando lo standard CSS, in questo modo grandi collezioni di dati vengono convertiti in oggetti SVG usando le funzioni di D3 e poi usati per generare rappresentazioni grafiche di numeri, testi, mappe e diagrammi.
I dati utilizzati possono essere in diversi formati, i più comuni sono JSON, CSV, TSV o geoJSON, ma, se necessario, si possono scrivere funzioni JavaScript apposite per leggere dati anche in altri formati.

D3 semplifica molto il lavoro dello sviluppatore tramite alcune funzionalità che non sono presenti nel linguaggio con cui è scritta, una di questa è il chaining che permette di concatenare l'uso di più funzioni applicate sullo stesso selettore, meccanismo ereditato (anche se non direttamente) da altre librerie come per esempio jQuery, alla quale somiglia anche per l'uso degli stessi selettori del CSS.

Il chaining risulta utile in quasi tutte le situazioni ma la più grande potenzialità di D3, e anche la sua più grande innovazione, consiste nel meccanismo del Data Binding.

Il Data Binding permette infatti allo sviluppatore di collegare i singoli dati ciascuno al proprio nodo del DOM. Per esempio, se si vuole tracciare uno Scatter Plot[5] sarà necessario disegnare molti cerchi che rappresentano le informazioni, utilizzando JavaScript il problema protrebbe essere risolto applicando un iterazione di questo tipo:

```
var dati = [0,2,3,4,5,6,7,8,9,]; // i nostri dati
for(var i=0;i<dati.length;i++){
    var svg = document.getElementsByTagName("svg")[0];
    var circle = document.createElement("circle"); // creo un elemento SVG cerchio
    circle.setAttribute("cx",Math.random()); //posizione coordinata X
    circle.setAttribute("cy",Math.random()); // posizione coordinata Y
    circle.setAttribute("r",Math.random()); // raggio del cerchio
    svg.appendChild(circle); //disegno il cerchio nella pagina
}
```

Figura 3.1: Creazioni di elementi svg con JavaScript

ovvero, creando un elemento cerchio con raggio uguale al dato che vogliamo visualizzare. Questo codice genera tanti cerchi, quanti gli elementi contenuti nell'array di partenza e ogni cerchio rappresenta un dato, ma che cosa accadrebbe se, in un secondo momento, i dati da visualizzare dovessero cambiare? Per esempio un'applicazione potrebbe richiedere, ad intervalli di tempo regolari, dei nuovi dati da un server remoto e le informazioni potrebbero essere più o meno numerose rispetto alla chiamata precedente. In questa situazione la funzione JavaScript dovrebbe prevedere delle funzionalità per cercare gli elementi da rimuovere, aggiornare i dati esistenti con i nuovi valori e infine aggiungere le nuove informazioni. Il risultato è che il codice potrebbe diventare molto articolato e quindi più soggetto ad errori, senza contare il tempo richiesto e la complessità elevata degli algoritmi da utilizzare.

D3 semplifica tutto questo proprio attraverso il Data Binding. I dati sono collegati direttamente ai nodi della pagina web e questo significa che se i dati dovessero cambiare, di conseguenza anche i nodi subirebbero delle modifiche. Il codice diventa molto più semplice e meno articolato. L'esempio precedente può essere scritto utilizzando D3 come nella figura 3.2.

Il codice è diviso in tre sezioni, la sezione Enter si occupa di inserire i nuovi dati che sono stati ricevuti, la sezione Exit si occupa di rimuovere i dati che non è più necessario visualizzare mentre la sezione Update si occupa di modificare i valori per i dati che continuano ad esistere ma che devono essere aggiornati.
Poiché i dati sono collegati direttamente ai rispettivi nodi la sezione Enter creerà un nuovo nodo ogni volta che viene ricevuto un nuovo dato, Exit rimuoverà i nodi per i dati che non esistono più e Update modificherà i parametri dei nodi con i valori aggiornati.

[5]Scatter Plot o grafico a dispersione è un tipo di grafico in cui due variabili di un set di dati sono riportate su uno spazio cartesiano.

```
var dati = [0,2,3,4,5,6,7,8,9,]; // i nostri dati
var svg = d3.select("svg"); //seleziona lo spazio SVG
var circle = svg.selectAll("circle")
    .data(dati);
var enter_circle = svg.enter().append("circle") //si disegnano i cerchi
    .attr("cx", Math.random()) //posizione coordinata X
    .attr("cy", Math.random()) //posizione coordinata Y
    .attr("r", function(d){return d}); //Raggio
var exit_circle = svg.exit().remove(); // rimuove i cerchi
circle.attr("cx", Math.random()) //posizione coordinata X
    .attr("cy", Math.random()) //posizione coordinata Y
    .attr("r", function(d){return d}); //Raggio
```

Figura 3.2: Creazioni di elementi svg con D3

D3 permette dunque la realizzazione di grafici e visualizzazioni molto complesse proprio grazie alla semplicità con il quale collega delle strutture dati al DOM di una pagina HTML, anche se questa tecnologia è relativamente nuova, la seconda versione di D3 è stata pubblicata dallo stesso Bostock pochi mesi dopo l'uscita della prima stesura, oggi siamo arrivati alla versione 3.5.14 e presto sarà disponibile la versione 4 dove, come annunciato dal suo sviluppare, oltre a contenere nuove funzioni e miglioramenti, sarà possibile utilizzare i singoli moduli che compongono la libreria piuttosto che importarla interamente.

3.2.3 Alcuni esempi di visualizzazioni

Visualizzare delle informazioni non è sempre facile e bisogna saper scegliere il giusto tipo di visualizzazione da applicare, un grafico ad esempio aiuta molto nella comprensione dei dati ma ne esistono di molti tipi e bisogna tenere conto dei loro pregi e difetti. Si deve inoltre considerare la percezione degli utenti quando osservano una visualizzazione e chiedersi quindi quale sarebbe la scelta migliore per quella tipologia di dati.

Un bar chart (grafico a barre) è un tipo di grafico molto diffuso e conosciuto. Questo grafico presenta i dati in maniera assoluta e viene infatti utilizzato sia per visualizzare dei singoli dati sia per presentare raggruppamenti di informazioni in modo da poter operare dei confronti.

Un esempio di visualizzazione non assoluta ma proporzionale è il pie chart (grafico a torta), questo grafico è composto da un cerchio diviso in tante parti quanti sono i dati da visualizzare, il rapporto tra una sezione e l'altra è proporzionale, ovvero indica un valore percentuale rispetto al totale del 100%. Il pie chart rende chiaro al primo sguardo il rapporto che intercorre tra i dati ma è spesso di difficile interpretazione poiché, non esprimendo un valore assoluto, può essere difficile stimare la quantità delle varie porzioni.

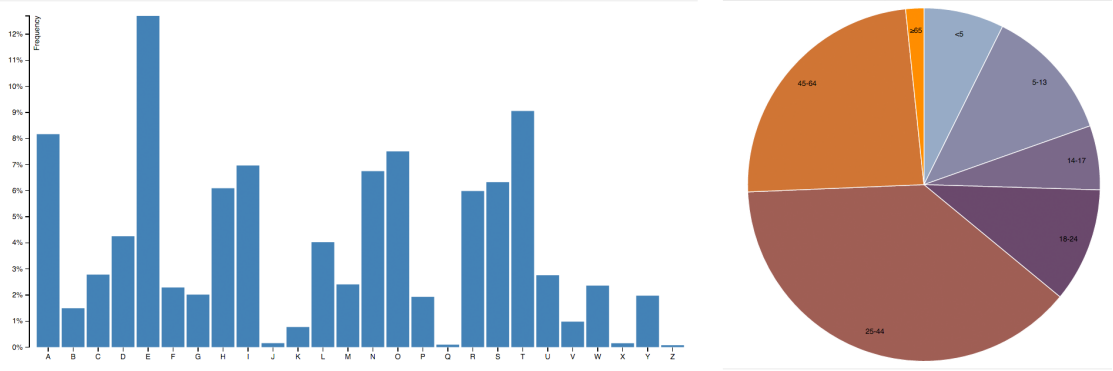


Figura 3.3: Bar chart e pie chart
http://bl.ocks.org/mbostock/3885304

Quasi sempre infatti ad ogni segmento si accosta il valore in percentuale della quantità espressa, purtroppo quando i dati sono molti o molto complessi il pie chart non riesce a rappresentarli al meglio ed è più efficace utilizzare un altro tipo di grafico.

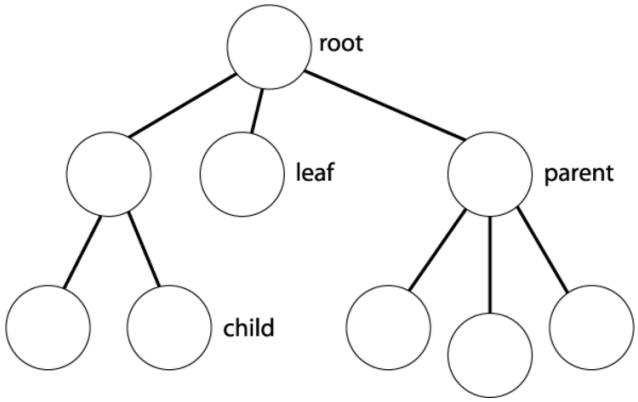


Figura 3.4: Visualizzazione di una struttura ad albero

Una visualizzazione un po' più complessa ma allo stesso tempo densa di informazioni è la treemap (mappa alberata), questo grafico è di tipo gerarchico, si basa infatti sulla visualizzazione di una struttura ad albero. In un albero il primo nodo è chiamato radice, i nodi collegati direttamente alla radice sono i suoi figli che a loro volta possono avere delle ramificazioni fino ad arrivare ai nodi terminali che sono chiamati foglie.

Nonostante una raffigurazione simile alla figura 3.4 sia un modo corretto di visualizzare un albero, non contiene però molte informazioni sui nodi stessi. Possiamo conoscere il tipo di dato che i nodi contengono e sicuramente le relazioni che intercorrono tra di essi, ma non abbiamo informazioni sul loro effettivo contenuto.

La treemap risolve questo problema visualizzando una struttura gerarchica e dando allo stesso tempo delle informazioni aggiuntive sui nodi. Ecco un esempio:

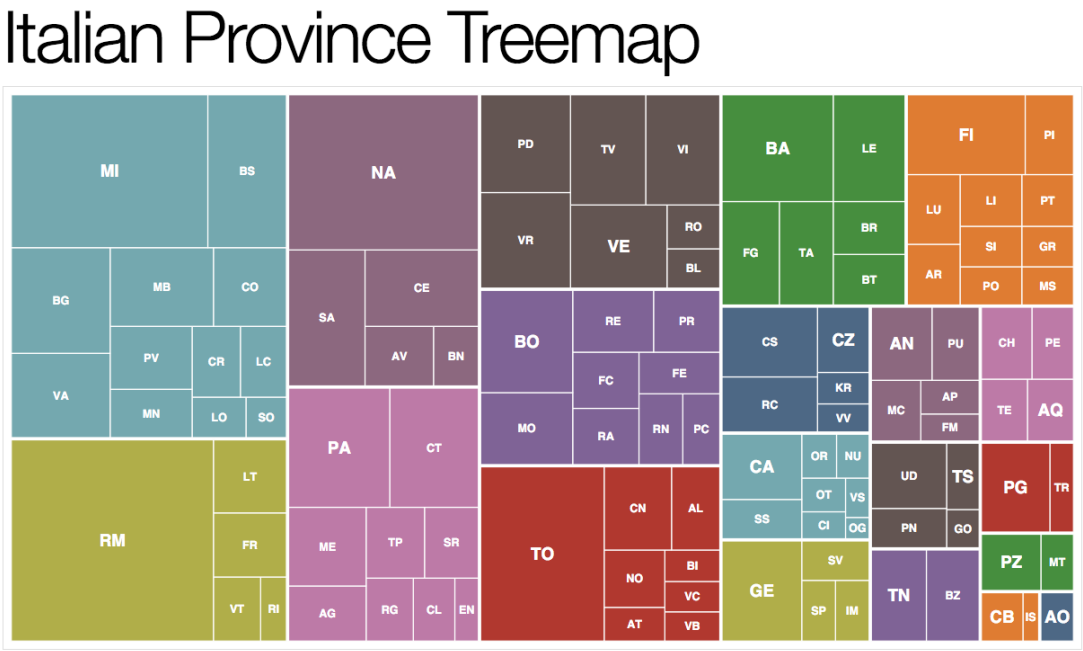


Figura 3.5: Treemap che visualizza la popolazione delle province e regioni italiane http://bl.ocks.org/fabiovalse/33baa8ba17fbdc83bdc6

Questa treemap mostra la popolazione delle regioni e province italiane. I rettangoli più grandi indicano la dimensione delle informazioni racchiuse nei nodi dell'albero, essi sono anche posizionati secondo la gerarchia originale della struttura. Il rettangolo in alto a sinistra è quindi il più grande e quello immediatamente sotto è il fratello diretto, all'interno dei rettangoli principali sono presenti altri rettangoli che indicano i figli e seguono la stessa gerarchia di dimensioni.

La struttura della treemap può essere complicata da comprendere in un primo momento ma, dopo le opportune spiegazioni sul metodo con il quale visualizza le informazioni, può diventare un valido strumento di visualizzazione.

Per capire quale grafico si presta meglio ad una serie di informazioni, si può operare un confronto tra la treeMap e uno stacked bar chart (grafico a barre aggregate) che mostra gli stessi dati.

Lo stacked bar chart segue la stessa struttura del bar chart con la differenza che ogni regione raggruppa una serie di dati e le varie provincie sono posizionate una affianco all'altra per permettere un confronto diretto. In questo caso si vede come nonostante entrambi i grafici rappresentino gli stessi dati, la treemap rappresenta meglio la struttura ad albero delle regioni italiane poiché rispetta la struttura gerarchica necessaria a

comprendere la relazione dei vari nodi[4].

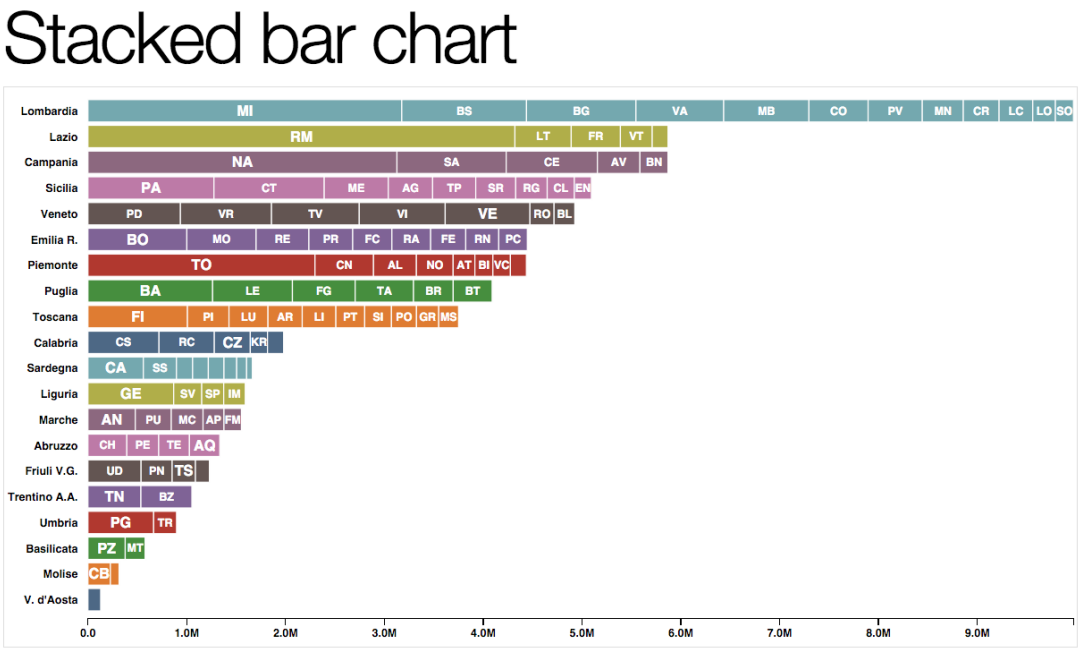


Figura 3.6: Stacked bar chart rappresentante le regioni e le province italiane http://bl.ocks.org/stefanomisuraca/3d5c2d3071e57ef7187b

3.3 Applicazioni basate su visualizzazioni: L'esperimento Meteo

I grafici mostrati precedentemente sono un ottimo esempio delle potenzialità della libreria D3, quelli presentati sono però solo grafici statici e non permettono nessun tipo di interazione con l'utente. Un grafico dinamico ha la capacità di modificare il modo in vengono visualizzati i suoi dati o semplicemente aggiornarsi con dei nuovi valori. Applicazioni di questo tipo sono molto importanti per visualizzare al meglio le informazioni di grandi dataset oppure per poter spostare l'attenzione da un gruppo di dati ad un altro.

Con l'aiuto della libreria D3 e della stazione meteorologica del CNR di Pisa è stata realizzata un'applicazione per la visualizzazione di dati come la temperatura, la radiazione solare, la forza del vento e molte altre informazioni geografiche e meteorologiche. L'applicazione consiste in un bar chart nel quale, in base al parametro scelto dall'utente, vengono disegnate le colonne che indicano il valore registrato dalla stazione nell'arco di un minuto. Vengono disegnate in tutto 180 colonne in modo da visualizzare un arco temporale di 3 ore di registrazione dei dati.

Per migliorare la visualizzazione le colonne non hanno spazi bianchi tra loro, così è più facile percepire la variazione dei dati nel tempo. Il grafico è inoltre interattivo poiché l'utente, tramite un drop-down menu, in qualsiasi momento può selezionare il tipo di dato da visualizzare, il risultato conseguente è l'aggiornamento delle colonne che cambieranno altezza per adattarsi ai nuovi valori.

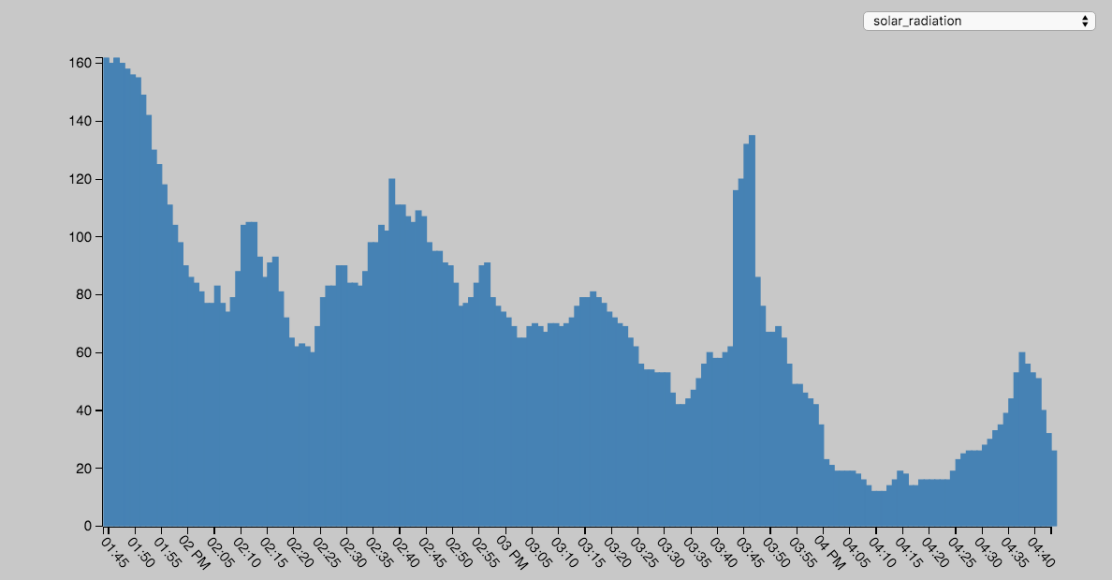


Figura 3.7: Esperimento Meteo: grafico a barre raffigurante la variazione della radiazione solare nell'arco di 3 ore

Una funzionalità molto interessante dell'applicazione è l'aggiornamento regolare dei dati. Ogni minuto la stazione meteorologica effettua una rilevazione di tutti i parametri e i dati vengono poi richiesti dall'applicazione al server, il programma aggiorna quindi i risultati e di conseguenza le colonne che contengono i valori. Queste ultime non cambiano la loro dimensione, piuttosto si spostano nella posizione appartenente al minuto precedente. Questo significa che ogni minuto tutte le colonne scalano verso sinistra poiché il dato che rappresentano è legato allo scorrere del tempo.

Questa caratteristica viene chiamata object constancy, ovvero costanza degli oggetti, questo perché una colonna, o più in generale un elemento della visualizzazione, mantiene costante il suo valore nel tempo senza mai alterarlo, ciò che si modifica è invece la posizione che via via porterà la colonna ad uscire fuori dall'intervallo temporale espresso nel grafico e quindi a non essere più visibile.

L'object constancy è molto importante per quanto riguarda il feedback che l'utente riceve, poiché quest'ultimo percepisce facilmente che i valori che sta osservando sono legati al momento nel quale li sta guardando, questa funzionalità contribuisce a dare ai dati un senso di coerenza assolvendo al compito principale dell'Information Visualization.

Avere un grafico interattivo e dinamico da la possibilità all'utente di analizzare con facilità una serie di dati che altrimenti sarebbe difficile da comprendere. D3, in particolare, aiuta lo sviluppo di applicazioni basate su visualizzazioni che sarebbero difficile da progettare e programmare utilizzando solo JavaScript o altri linguaggi.

Durante la fase di progettazione bisogna sempre ricordare le tre domande a cui una visualizzazione deve dare risposta e cercare di capire se un grafico interattivo può aiutare e in che misura l'analisi del contenuto della visualizzazione. A volte può essere opportuno creare un grafico statico, altre invece la grande mole di dati e loro complessità rende necessario creare dei grafici dinamici, il progettista deve dunque sempre tenere in considerazione lo scopo della visualizzazione e l'utenza a cui è rivolta.

Capitolo 4

Il Replayer: una soluzione al problema visualizzazione

4.1 Introduzione

Il Replayer è un applicazione web che sfrutta tecniche di social sensing e di visualizzazione dell'informazione per mostrare i dati raccolti dagli utenti in reazione ad un evento significativo che si svolge lungo un arco temporale definito.

L'applicazione è composta da tre sezioni principali: una mappa, che mostra la localizzazione geografica sia del luogo dell'evento sia dei dati raccolti, una word cloud (o tag cloud), cioè un tipo di visualizzazione che rappresenta i tag o le parole chiave estratte dai dati e le ordina secondo dei criteri predefiniti come la frequenza o la popolarità e infine un player dedicato al controllo dell'applicazione in modo che l'utente possa interagire con essa in maniera dinamica per filtrare, approfondire o analizzare le informazioni presentate.

Il Replayer prende in esame eventi realmente accaduti, sono stati selezionati avvenimenti recenti e di interesse europeo o italiano come la tromba d'aria nei pressi di Venezia o gli attentati di Parigi del 13 Novembre 2015. In queste occasioni gli utenti del Web sono stati i primi a commentare e diffondere le notizie sui social network. Estrapolando queste informazioni, il Replayer colloca sulla mappa tutti i post e i tweet contenenti informazioni, foto o commenti esattamente nel punto e nel momento in cui sono stati scritti.

4.1.1 Dati utilizzati

Le persone utilizzano i social network ogni giorno e li sfruttano per pubblicare principalmente informazioni riguardo le loro esperienze personali e ciò che gli accade intorno, quando si verifica un evento sociale di particolare interesse gli utenti condividono, spesso in maniera istantanea, le informazioni in loro possesso scrivendo dei messaggi oppure scattando delle foto.

Gli utenti, con i loro post, tweet e foto, fungono da monitor sociale e forniscono uno specchio della situazione praticamente in tempo reale, il Replayer sfrutta proprio queste informazioni per visualizzare sulla mappa i dati raccolti e ampliare la conoscenza in merito ad un evento ancora nel corso del suo svolgimento.

È necessario fare però una precisazione riguardo i dati, poiché è doveroso tutelare la privacy degli utenti e non è possibile prelevare dei dati sensibili senza il permesso delle persone coinvolte, il Replayer si basa su metadati, ovvero informazioni parziali che il social network in questione permette di estrarre e che solo in seguito vengono analizzate e arricchite con l'aggiunta di stime e probabilità calcolare in fase di esecuzione.

I social network presi in considerazione sono: Facebook, Twitter e Instagram. In particolare quest'ultimo è stato scelto per la possibilità di visualizzare le foto dei post degli utenti. Tutti questi dati risultano di fondamentale importanza, non solo per comprendere al meglio l'evento con delle testimonianze dirette di chi lo sta vivendo in prima persona, ma anche per avere delle informazioni significative per le forze dell'ordine o le unità di soccorso per agevolare le loro indagini o per risolvere al meglio una situazione di emergenza.

Figura 4.1: Replayer: tromba d'aria in Veneto

4.1.2 Finalità

Lo scopo dell'applicazione è visualizzare in maniera dinamica le reazioni degli utenti di fronte a eventi di grosso calibro, si possono misurare infatti sia la tempestività di reazione nello scrivere messaggi e scattare foto sia la diffusione della notizia e quindi il conseguente incremento di post e tweet con il passare del tempo.

L'applicazione aiuta dunque ad analizzare i messaggi scritti, il contenuto delle foto scattate e le parole più utilizzate, queste ultime in particolare sono molto utili per capire come si sviluppano fenomeni sociali come gli hashtag, ovvero delle parole chiave che identificano un concetto condiviso da molte persone, la loro funzione è di rendere più facile per gli utenti trovare messaggi riguardanti un tema o un contenuto specifico.
Un utente può creare o utilizzare un hashtag posizionando il carattere hash (cancelletto #) davanti ad una parola o una frase del testo di un messaggio, successivamente la ricerca di quel hashtag restituirà tutti i messaggi che sono stati etichettati con esso, per avere un esempio concreto basti pensare al famoso hastag #PrayForParis che diventò famoso durante gli attentati a Parigi.

Questa applicazione fornisce quindi molti spunti sia per analisi sociologiche o statistiche riguardo i meccanismi di reazione delle persone. Permette inoltre di aumentare la nostra conoscenza riguardo un evento di interesse nazionale o mondiale tramite una visualizzazione interattiva che fornisce una visione di insieme sul fatto esaminato aggregando tutte le informazioni disponibili tratte dai social media e si prefigura come un valido strumento per la gestione delle emergenze e l'agevolazione dei soccorsi e delle forze dell'ordine.

4.2 stato dell'arte

Per la realizzazione del Replayer sono stati necessari molti studi preliminari per approfondire concetti e problematiche di visualizzazione dell'informazione e linguistica computazionale. È stato dunque molto importante identificare il metodo migliore per strutturare un'applicazione che dovesse gestire una quantità di dati molto elevata e sopratutto visualizzarli in maniera efficace.

Il Replayer nasce dall'idea che gli eventi sociali, sia essi positivi come un concerto o una manifestazione, sia negativi come una calamità naturale o una situazione di emergenza, sono caratterizzati da un feedback molto forte da parte degli utenti del web. I commenti, le foto e i messaggi che gli utenti si scambiano riguardo l'evento sono una testimonianza diretta dell'accaduto, queste informazioni sono il fulcro centrale dei social media.

Sfruttando queste informazioni si possono estrarre notizie o creare applicazioni di social monitoring e social sensing, un esempio è SoS - Social Sensing, progetto del CNR di Pisa, che ha sfruttato i dati estratti dai social media per approfondire e studiare il terremoto in Emilia del 2012.
Così facendo quest'applicazione di social monitoring ha contribuito a presentare l'accaduto dal punto di vista degli utenti che sono stati colpiti dal terremoto. Proprio da questo progetto e prendendo spunto da altri strumenti di social sensing nasce il Replayer.

Lo scopo è quello di rendere l'applicazione per il social monitoring molto versatile e potente ma utilizzabile anche come strumento di urban uensing, cioè un modo per studiare le dinamiche sociali delle città. Lo scopo principale è la prevenzione delle emergenze, ovvero identificare gli eventi sociali potenzialmente dannosi e contattare le forze dell'ordine o allertare i cittadini prima che si verifichino danni consistenti.

4.3 Grafici utilizzati

4.3.1 Bar chart

Per presentare la crescita dei dati con il passare del tempo dall'inzio dell'evento preso in esame è stato scelto un bar chart, è possibile quindi a colpo d'occhio identificare la frequenza delle informazioni e la loro propagazione, così facendo l'utente percepisce subito l'andamento della quantità di post e tweet inseriti dagli utenti.

Data la grande quantità di informazioni si è optato per un raggruppamento dei dati, che sarebbero stati illeggibili se presentati singolarmente, in questo modo la distinzione tra il periodo di tempo prima dello scatenarsi dell'evento è subito identificabile dalla poca presenza di informazioni, man mano che la notizia si diffonde appaiono nel grafico dei picchi di informazioni, inoltre un contatore posto sotto di esso fornisce il numero esatto dei dati che si stanno visualizzando.

Per migliorare la comprensione e l'interazione da parte dell'utente il grafico è stato dotato di un cursore che scorre al passare del tempo in modo da percepire la dimensione temporale dell'avvenimento, questo marcatore temporale è interattivo e può essere spostato dall'utente per analizzare un preciso momento dello svolgimento dell'evento. Integrato al grafico vi è il player che permette di controllare la riproduzione dell'applicazione, non solo dandole inizio o fermandola, ma anche aumentando la velocità di presentazione delle informazioni. È inoltre sempre presente un orologio che in indica la data e l'ora dei fatti che si stanno osservando.

4.3.2 Visualizzazioni: word cloud

La word cloud è una visualizzazione che permette di identificare, all'interno di un insieme di parole, quelle con il valore più alto rispetto al criterio di ordinamento scelto, nel caso del Replayer si tratta della frequenza di utilizzo all'interno dei tweet e post raccolti. Di conseguenza le parole che hanno una dimensione maggiore sono anche le più utilizzate. La struttura di una word cloud tende ad essere confusionaria perché in un vasto insieme di parole è difficile identificare la corretta scala di valori e tra due oggetti che hanno circa la stessa frequenza risulta difficile distinguere quello con un valore o una dimensione maggiore.

In una visualizzazione di questo tipo è immediato identificare le parole che hanno un valore maggiore poiché esse avendo una dimensione più grande risultano facilmente visibili. Al diminuire della frequenza diminuisce però anche la capacità di percepire la variazione della grandezza fra le parole, quindi questo strumento di visualizzazione non è indicato per fornire una panoramica completa su tutti i dati, quanto più per identificare solo quelli di maggiore importanza.

Nel caso del Replayer però è stata scelta una word cloud per permettere all'utente di osservare tutte le parole che sono state estratte dai social media e dato il grande cambiamento della frequenza dei dati e l'aggiunta di nuove parole con il passare del tempo, questa visualizzazione dinamica si presta bene per la presentazione di questo tipo di informazioni poiché l'utente può identificare con facilità le parole che stanno

Figura 4.2: Esempio di word cloud, le parole sono ordinate sia per dimensioni che per colore

acquistando una maggiore popolarità nello svolgersi dell'evento. Inoltre cliccando su una parola essa viene evidenziata così che chi utilizza l'applicazione può focalizzarsi su una singola parola e vederne l'andamento nel tempo.

4.3.3 Mappa

Per la realizzazione della mappa sono state prese in considerazione molte librerie JavaScript tra cui Google Maps, OpenLayers e MapBox, queste soluzioni sono sembrate valide in un primo momento ma in seguito sono emersi dei difetti per quanto riguarda la libertà di utilizzo. In particolare Google Maps e MapBox non consentivano un utilizzo intensivo dell'applicazione, limitando per esempio le chiamate alle API (Application Programming Interface) oppure veniva richiesto di condividere le informazioni a fini statistici mettendo a rischio la privacy e l'utilizzo commerciale dei dati.

Per poter aggirare questi problemi è stata scelta Leaflet[1], ovvero una libreria JavaScript che permette la creazione di una mappa senza particolari limitazioni. Questa libreria mette a disposizione uno strumento molto versatile e semplice da usare, è infatti possibile creare marker (segnaposto) o qualsiasi tipo di figure sulla mappa semplicemente indicando le coordinate o le misure esatte. Inoltre la mappa generata non richiede molte risorse per l'elaborazione e questo ha agevolato il posizionamento di molti marker, anche nell'ordine di migliaia, che vanno ad indicare il punto esatto nel quale un tweet o un post è stato inviato da parte dell'utente. Il risultato è un agglomerato di marcatori che mostrano da quali zone, all'interno dell'area di interesse dell'evento sociale, provengono i messaggi degli utenti.

[1] http://leafletjs.com/

La mappa è interattiva, infatti l'utente può utilizzare le funzioni di zoom per osservare meglio una data zona, inoltre il diverso colore dei marcatori ha reso possibile la creazione un sistema di codifica per distinguere il social network di provenienza dell'informazione. L'utente può interagire con la mappa anche cliccando sui marker, così facendo viene aperto un link che riporta alla risorsa originale da cui sono state tratte le informazioni, quindi i tweet, post su Facebook oppure foto tratte da Instragram.

4.4 Sviluppo e funzionalità

4.4.1 Paradigma MVC

La progettazione di un'applicazione molto articolata come il Replayer può essere molto complessa, di solito le applicazioni web vengono scritte secondo un paradigma di programmazione procedurale, il che significa che il codice è composto da una serie di funzioni o procedure che vengono eseguite una di seguito all'altra formando appunto una procedura. Quando un'applicazione richiede la scrittura di un codice molto esteso però, questo paradigma può non essere molto efficace, non tanto per la sua complessità computazionale ma dal punto di vista dello sviluppatore che, dovendo scrivere anche migliaia di righe di codice, potrebbe trovare delle difficoltà nella stesura o nel comprendere script scritti da altri sviluppatori che collaborano al progetto.

Per questo motivo, sopratutto nell'ambito di progetti complessi, si preferisce utilizzare il paradigma di programmazione orientato agli oggetti (OOP), pratica in uso già nei primi anni settanta ma che è diventata dominante negli anni novanta dopo l'introduzione di alcune versioni ad oggetti dei linguaggi di programmazioni più comuni, in particolare il C++.

Questo tipo di programmazione trova il suo punto di forza nella definizione di oggetti software che possono interagire gli uni con gli altri, questi oggetti fanno parte di una struttura più grande che li raggruppa, secondo i criteri scelti dal programmatore, in classi. Una classe è un modello nel quale si definiscono tutte le proprietà ed i metodi delle strutture che appartengono alla classe stessa, successivamente questi oggetti vengono istanziati durante l'esecuzione del programma seguendo sempre le direttive della classe di appartenenza. Questo metodo di programmazione risulta molto efficace nella creazione di interfacce grafiche, proprio per questo il Replayer, poiché è un'applicazione che sfrutta un'interfaccia direttamente controllata dagli utenti, è stato scritto secondo il paradigma ad oggetti.

Nonostante questo metodo di programmazione sia molto efficace spesso può risultare più articolato poiché bisogna saper progettare bene le classi, le eventuali sottoclassi e saper sfruttare le caratteristiche principali di tale paradigma come l'ereditarietà o l'overloading.

Per aiutare i programmatori a sviluppare progetti di una certa complessità è stato studiato MVC (Model-View-Controller), un pattern architetturale molto diffuso nello sviluppo di sistemi software, in particolare nell'ambito della programmazione orientata agli oggetti. Storicamente il pattern MVC è stato implementato lato server ma recente-

mente, con lo sviluppo e la parziale standardizzazione di JavaScript sono nate le prime implementazioni lato client.

Negli ultimi anni è aumentata la richiesta di applicazioni Web basate su chiamate asincrone al server utilizzando la tecnologia AJAX (Asynchronous JavaScript And XML), che permette di visualizzare i risultati delle elaborazioni senza aggiornare la pagina o eseguire redirect. Con la diffusione di JavaScript si è sentita l'esigenza di creare i primi framework che implementino il paradigma MVC in questo linguaggio, uno dei primi è stato Backbone.js, seguito da una serie interminabile di altri framework, tra cui JavaScriptMVC, Ember ed AngularJS.
Il pattern è basato sulla separazione dei compiti fra i componenti software che interpretano tre ruoli principali:
il model fornisce i metodi per accedere ai dati utili all'applicazione. Il view visualizza i dati contenuti nel model e si occupa dell'interazione con gli utenti, il controller riceve i comandi dell'utente, in genere attraverso il view, e li attua modificando lo stato degli altri due componenti. MVC è comunque una logica non perfettamente definita e l'implementazione di tale architettura dipende strettamente dall'applicazione e dalle tecnologie che si stanno utilizzando per programmare. Durante lo sviluppo di Replayer è stato utilizzata la logica di MVC che però è stata leggermente modificata per adattarsi al meglio all'applicazione.[8]

4.4.2 Complessità e difficoltà

Lo sviluppo del Replayer è durato diversi mesi, il team di sviluppo ha lavorato costantemente alla produzione degli script necessari per il suo funzionamento, come in tutti i progetti però sono state incontrate delle difficoltà e delle sfide nel trovare la soluzione migliore ai problemi che si sono presentati. Lo sviluppo della mappa per esempio ha richiesto uno studio approfondito sulla scelta dei tiles da utilizzare, la libreria Leaflet mette a disposizione uno strumento per disegnare sulla mappa ma essa non offre la mappa geografica vera e propria, è stato dunque necessario ricercare un servizio che mettesse a disposizione il proprio tile server. Per fare questo ci siamo affidati a OpenStreetMap, servizio che concede l'utilizzo dei propri tile server senza troppe restrizioni.

Un altro problema con il quale il team di sviluppo si è trovato a confronto è stato trovare un modo per poter visualizzare al meglio la word cloud. Questa purtroppo, con l'avanzare del cursore del player attraverso l'asse temporale, cresce sempre di più in numero e grandezza delle parole e non è più possibile distinguere bene gli elementi. Per ovviare a questo problema si è scelto di utilizzare la word cloud ma di organizzarla secondo il layout di una treemap. In questo modo tutte le parole che vengono visualizzare hanno una loro specifica posizione che rende facile l'identificazione di una parola e sopratutto la sua posizione all'interno della scala di valori.

4.5 Sviluppi futuri e utilizzi possibili

Gli utilizzi possibili del Replayer sono molteplici, in particolar modo è stato pensato come uno strumento che agevoli le forze dell'ordine nella prevenzione e nella gestione delle emergenze, infatti raggruppando tutte le informazioni raccolte sui social network si può avere una visione complessiva e immediata di quello che sta accadendo in un dato luogo prima ancora delle notizie ufficiali rilasciate dai media tradizionali infatti il divario temporale tra il verificarsi di un evento e la comparsa delle prime notizie è molto inferiore nei social network poiché utilizzano come fonti proprio le persone coinvolte nell'avvenimento mentre nei media tradizionali possono passare anche ore prima della diffusione delle notizie.

Lo sviluppo futuro del progetto è l'implementazione di nuove funzionalità al fine di rendere il Replayer un vero e proprio social monitor live, cioè sviluppare delle tecnologie che gli consentano di identificare l'inizio di un evento di interesse sociale e iniziare quindi a registrare in autonomia tutti i dati che gli utenti inseriscono in merito a questo fenomeno, così facendo il Replayer potrebbe individuare quali sono gli avvenimenti che generano maggior interesse all'interno del Web e compiere delle analisi sui tweet, post e informazioni inserite dagli utenti.

Per quanto riguarda gli strumenti di visualizzazione utilizzati, in futuro si potranno migliorare e aggiungerne di nuovi, un line chart potrebbe essere integrato al di sopra del bar chart al fine di presentare nuove informazioni oltre alla quantità dei dati, ad esempio l'incremento del numero di persone che vengono a conoscenza dell'evento tramite i social media, sono già stati realizzati dei prototipi per l'integrazione di questi grafici.
Inoltre saranno fatti ulteriori esperimenti per trovare nuove metodologie per la presentazione dai dati, infatti per alcuni avvenimenti è stato preferito mostrare la cronologia dei tweet mentre per altri la word cloud con i tag più popolari, sarà dunque necessario identificare le visulizzazioni migliori a seconda della portata dell'evento.

Si potranno inoltre integrare altri social media popolari per avere a disposizione un numero di informazioni sempre maggiori, ad esempio YouReporter una piattaforma di citizen journalism dove gli utenti possono creare video, foto e servizi per raccontare la realtà che li circonda, inoltre il Replayer dovrà essere pensato per dare spazio anche a tutti i social media che saranno sviluppati in futuro ed essere sempre pronto a stare al passo con il rapido movimento del Web.

Tutti questi miglioramenti potrebbero diventare realtà in un futuro non molto lontano. Ciò che aiuterà il Replayer ad espandersi e migliorarsi sempre di più è un controllo preciso sul codice scritto, l'architettura MVC ha infatti aiutato molto lo sviluppo e il riutilizzo del codice, si pensa tuttavia in futuro di passare da un modello MVC pensato e generato in autonomia ad una struttura di MVC basata su un framework JavaScript al fine di migliorare la stabilità ed avere un supporto da parte dei programmatori che fanno parte della community di sviluppo del framework stesso.

In particolare è stato scelto il framework MVC Backbone.js che permetterà un maggiore controllo e riutilizzo del codice, il framework tuttavia limita un po' lo sviluppo in quanto bisogna rispettare le specifiche dettate da questa tecnologia al fine di garantire

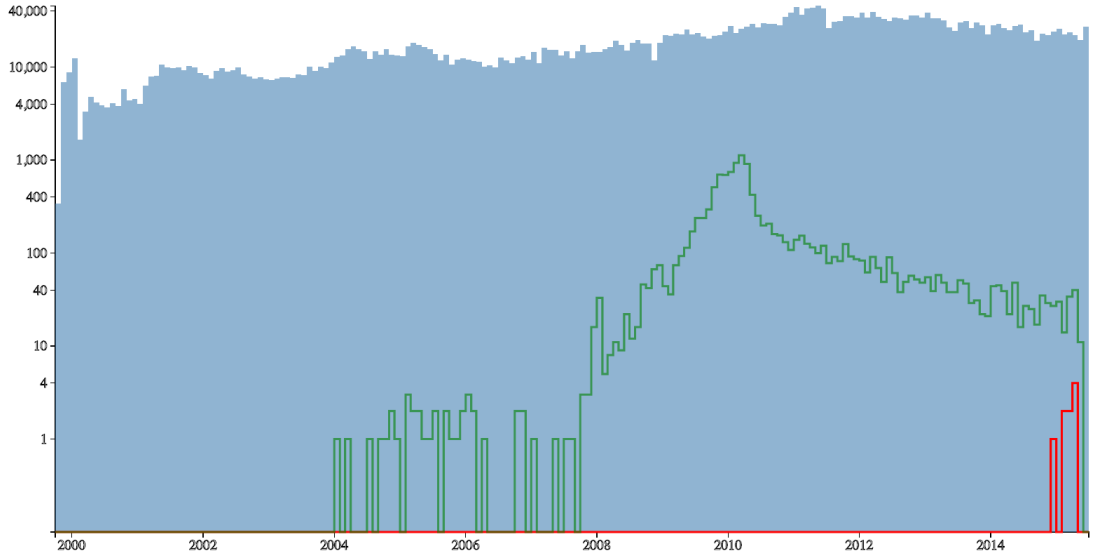


Figura 4.3: Prototipo di due line chart sovrapposti ad un bar chart che visualizzano dati diversi

un miglior risultato finale, bisogna dunque valutare attentamente le conseguenze delle scelte future poiché in una applicazione articolata come il Replayer bisogna tenere in considerazione sia i benefici di una struttura più rigida che le limitazioni.

Conclusioni

Il lavoro di ricerca alla base di questa tesi ha reso possibile l'approfondimento di tecniche di visualizzazione per presentare un grande quantitativo di dati tramite nuove tecnologie e linguaggi come la libreria Data Driven Documents e l'utilizzo di grafici dinamici e interattivi. Per comprendere meglio il contesto di lavoro sono state condotte delle ricerche sui social media, i big data e le tecniche più recenti di presentazione delle informazioni.

Queste ricerche hanno portato ad un contributo nello sviluppo dell'applicazione Replayer che si propone di essere uno strumento di social sensing per l'analisi di eventi di interesse nazionale e internazionale tramite l'elaborazione dei dati ricavati dai social media più diffusi come Twitter, Facebook e Instagram. In particolare l'applicazione si configura come un monitor sociale per la gestione delle emergenze e uno strumento per agevolare le forze dell'ordine nel condurre le indagini e reperire in modo più veloce ed efficace le informazioni riguardanti un evento ancora in corso.

Gli eventi campione analizzati includono gli attentati di Parigi 2015 e le alluvioni in Veneto, tramite lo studio dei dati raccolti sono state ideate tecniche di visualizzazione per presentare le informazioni tra cui una mappa interattiva, una word cloud e un bar chart. Utilizzando l'applicazione gli utenti possono navigare ed esplorare le informazioni per ampliare la loro conoscenza in merito all'evento in oggetto e avere accesso ad un quantitativo di dati che sarebbero incomprensibili se non opportunamente visualizzati.

Lo sviluppo futuro del Replayer include l'aggiunta di nuove funzionalità, in particolare una modalità live che reperisca ed elabori le informazioni dei social media in tempo reale in modo da fornire uno specchio della realtà in continuo aggiornamento.

Bibliografia

[1] Alessandro Acquisti and Ralph Gross. Imagined communities: Awareness, information sharing, and privacy on the facebook. In *Privacy enhancing technologies*, pages 36–58. Springer, 2006.

[2] Charu C Aggarwal. *Managing and mining sensor data*. Springer Science & Business Media, 2013.

[3] Marco Avvenuti, Stefano Cresci, Andrea Marchetti, Carlo Meletti, and Maurizio Tesconi. Ears (earthquake alert and report system): A real time decision support system for earthquake crisis management. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, pages 1749–1758, New York, NY, USA, 2014. ACM.

[4] Christian Behrens. *The Form of Facts and Figures*. 2008.

[5] Tim Berners-Lee. *L'architettura del nuovo Web. Dall'inventore della rete il progetto di una comunicazione democratica, interattiva e intercreativa*. Feltrinelli Editore, 2001.

[6] Tim Berners-Lee, James Hendler, Ora Lassila, et al. The semantic web. *Scientific american*, 284(5):28–37, 2001.

[7] Stuart K Card and Jock Mackinlay. The structure of the information visualization design space. In *Information Visualization, 1997. Proceedings., IEEE Symposium on*, pages 92–99. IEEE, 1997.

[8] Alexander Chatzigeorgiou and George Stephanides. *Evaluating performance and power of object-oriented vs. procedural programming in embedded processors*. Springer, 2002.

[9] World Wide Web Consortium et al. Extensible markup language (xml) 1.1. 2006.

[10] Richard L Daft and Robert H Lengel. Organizational information requirements, media richness and structural design. *Management science*, 32(5):554–571, 1986.

[11] Nicole B Ellison et al. Social network sites: Definition, history, and scholarship. *Journal of Computer-Mediated Communication*, 13(1):210–230, 2007.

[12] Roy Fielding, Jim Gettys, Jeffrey Mogul, Henrik Frystyk, Larry Masinter, Paul Leach, and Tim Berners-Lee. Hypertext transfer protocol–http/1.1. Technical report, 1999.

[13] Raghu K Ganti, Nam Pham, Yu-En Tsai, and Tarek F Abdelzaher. Poolview: stream privacy for grassroots participatory sensing. In *Proceedings of the 6th ACM conference on Embedded network sensor systems*, pages 281–294. ACM, 2008.

[14] David E Goldberg and John H Holland. Genetic algorithms and machine learning. *Machine learning*, 3(2):95–99, 1988.

[15] Charlotte N Gunawardena. Social presence theory and implications for interaction and collaborative learning in computer conferences. *International journal of educational telecommunications*, 1(2):147–166, 1995.

[16] Hanqi Guo, Ningyu Mao, and Xiaoru Yuan. Wysiwyg (what you see is what you get) volume visualization. *Visualization and Computer Graphics, IEEE Transactions on*, 17(12):2106–2114, 2011.

[17] Andreas M Kaplan and Michael Haenlein. Users of the world, unite! the challenges and opportunities of social media. *Business horizons*, 53(1):59–68, 2010.

[18] Julia Layton. Amazon technology. *Money. howstuffworks. com*, 2013.

[19] Leah A Lievrouw and Sonia Livingstone. *Handbook of new media: Social shaping and consequences of ICTs*. Sage, 2002.

[20] J Lynn. Internet users to exceed 2 billion this year. *Retrieved*, 3(14):2011, 2010.

[21] James Manyika, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh, and Angela H Byers. Big data: The next frontier for innovation, competition, and productivity. 2011.

[22] Simone Martini and Maurizio Gabbrielli. *Linguaggi di programmazione-Principi e paradigmi*. McGraw-Hill, Italia,, 2006.

[23] Michele Mauri and Paolo Ciuccarelli. 5. il ruolo dell'information visualization nella progettazione di interfacce per archivi digitali eterogenei. *Quaderni DigiLab*, 3(1):73–88, 2014.

[24] Tamara Munzner. *Visualization Analysis and Design*. CRC Press, 2014.

[25] Office of Science and Technology Policy Executive Office of the President. Obama administration unveils "big data" initiative: Announces 200 million in new r&d investments, mar 2012.

[26] Tim O'reilly. What is web 2.0: Design patterns and business models for the next generation of software. *Communications & strategies*, (1):17, 2007.

[27] Alessandro Prunesti. Statistiche e trend su internet, social media e mobile per il 2014 in italia e nel mondo, jan 2014.

[28] Ulf-Dietrich Reips and Jochen Musch. A brief history of web experimenting. *Psychological experiments on the Internet*, page 61, 2000.

[29] Arthur L Samuel. Some studies in machine learning using the game of checkers. *IBM Journal of research and development*, 3(3):210–229, 1959.

[30] Bruce R Schatz and Joseph B Hardin. Ncsa mosaic and the world wide web: global hypermedia protocols for the internet. *SCIENCE-NEW YORK THEN WASHINGTON-*, pages 895–895, 1994.

[31] Oreste Signore. Introduzione al semantic web, may 2008.

[32] Dong Wang, Tarek Abdelzaher, and Lance Kaplan. *Social sensing: building reliable systems on unreliable data*. Morgan Kaufmann, 2015.

[33] C. Ware. *Information Visualization: Perception for Design*. Morgan Kaufmann (Academic Press), 2000.