Tomotopigrafie

Modelli visivi per processi di topic modeling dinamico e gerarchico

Una nota sul testo
Leitura, Dino Dos Santos, 2007, utilizzato nelle versioni Roman1 e Italic1.
Aktiv Grotesk, Dalton Maag, 2010, utilizzato in Regular, Medium e nelle relative versioni italiche.

Acroprint extra white, 120gr
Satogami light green, 116gr

Stampato a Milano nel mese di Aprile 2018 presso Graphic S.R.L. e SEF

Rilegato a Brugherio (MB) presso Arte Libro S.N.C.

POLITECNICO
MILANO 1863

Tesi di Beatrice Gobbo
Matricola 840793
Relatore: Paolo Ciuccarelli
Correlatore: Michele Mauri

Politecnico di Milano
Scuola del Design
LM — Design della Comunicazione A.A. 2016/2017

La prima regola era di non accettare mai nulla per vero, senza conoscerlo evidentemente come tale: cioè di evitare scrupolosamente la precipitazione e la prevenzione; e di non comprendere nei miei giudizi niente più di quanto si fosse presentato alla mia ragione tanto chiaramente e distintamente da non lasciarmi nessuna occasione di dubitarne.
La seconda, di dividere ogni problema preso in esame in tante parti quanto fosse possibile e richiesto per risolverlo più agevolmente.
La terza, di condurre ordinatamente i miei pensieri cominciando dalle cose più semplici e più facili a conoscersi, per salire a poco a poco, come per gradi, sino alla conoscenza delle più complesse; supponendo altresì un ordine tra quelle che non si precedono naturalmente l'un l'altra.
E l'ultima, di fare in tutti i casi enumerazioni tanto perfette e rassegne tanto complete, da essere sicuro di non omettere nulla.

— *René Descartes, Discorso sul Metodo, 1637*

Abstract

Nell'era dell'*information overload* l'utente ha necessità di trovare informazioni specifiche senza dover consultare immensi archivi documentali, digitali e non. Il *topic modeling* è un modello statistico basato sull'uso di una serie variabile di algoritmi che, a partire da un esteso gruppo di documenti (*corpus*), permette di identificare gli argomenti (topic) trattati a partire dall'identificazione delle parole chiave (*keywords*). Nell'ultimo decennio il contributo della data visualization è diventato fondamentale per migliorare l'analisi e la fruizione dei dati. Rivolgendosi da sempre ad un pubblico esperto e di nicchia la visualizzazione dei risultati di *topic modeling* è sempre stata legata a modelli visivi sedimentati e noti come lo *streamgraph*, le *word clouds* e il *force directed graph*. Durante la fase di ricerca iniziale di questa tesi, è emerso come un approccio metaforico alla complessità dei dati possa agevolarne la rappresentazione e la fruibilità attraverso un'interfaccia semplice e intuitiva.
Inizialmente in *Topic Tomographies* (*TopTom*), progetto nato in collaborazione con ISI Foundation, è stato sperimentato l'uso di modelli visivi noti che evocassero tecniche di rappresentazione del corpo umano provenienti dall'ambito medico. Infine, le criticità e i limiti del progetto hanno consentito di analizzare a fondo caratteristiche e problematiche dei dati, portando in primo piano l'interesse per la definizione e la visualizzazione del topic in sé, entità astratta e generata automaticamente da un algoritmo.

1. Introduzione

1.1 Premessa

In un periodo storico-sociale in cui le informazioni ricavate dal web sotto forma di dati rappresentano una miniera d'oro per mappare la società, gli algoritmi di *machine learning* sono sempre più diffusi ed utilizzati per definire struttura ed evoluzione dei contenuti nelle discussioni online senza dover consultare immensi archivi di documenti. È infatti quasi riduttivo pensare che l'output di calcoli algoritmici siano solo gli ormai famosissimi suggerimenti di Netflix, il ranking del Facebook News Feed o le playlist personalizzate su Spotify.

> *"An algorithm is a recipe, an instruction set, a sequence of tasks to achieve a particular calculation or result, like the steps needed to calculate a squre root or tabulate the Fibonacci sequence"*
>
> — *E. Finn, 2017*[1]

Un algoritmo è una ricetta, un set di istruzioni, una

1. Finn, E., What Algorithms Want - Imagination in the Age of Computing p.17 , MIT Press, (2017)

serie di compiti necessari per raggiungere un determinato risultato di calcolo.
I *topic models* sono un insieme di modelli statistici basati sull'uso di *algoritmi supervisionati* e *non supervisionati* [2] che permettono di identificare topic (gruppi di parole che definiscono un argomento) a partire da un esteso gruppo di documenti. Negli ultimi dieci anni le ricerche relative a questo settore sono cresciute in maniera esponenziale, rendendo sempre più complessa la forma dei risultati prodotti.
Negli ultimi dieci anni, gli algoritmi di *topic modeling* sono stati migliorati e implementati e, negli ultimi due anni, la necessità di rappresentare risultati sempre più complessi e annidati ha dato inizio alla collaborazione tra il mondo della *data science* e quello del design dell'informazione. Già partire dalle prime sperimentazioni, la visualizzazione del *topic modeling* si è sempre affidata a modelli visivi noti come lo *streamgraph*[3] e il *force directed graph*[4] . Ad oggi, la letteratura non sembra ancora offrire un esempio di strumento che, identificando un modello visivo di partenza attraverso l'uso di metafore, permetta di navigare in maniera dinamica tra le viste, soprattutto in presenza di risultati estratti con *topic modeling* sia dinamico che gerarchico (cap.☞ 3).
Il progetto *TopTom - Topic Tomographies* è stato punto di partenza e spunto di ricerca per *Tomotopigrafie. Modelli visivi per processi di topic modeling dinamico e gerarchico*. Nella ricerca si mostrerà come un approccio metaforico e multi-vista, basato sull'uso di modelli visivi dinamici, possa migliorare l'esperienza utente e come la rappresentazione del comportamento del topic attraverso uno specifico modello visivo possa mettere in luce pattern di comportamento che non dipendono dal contenuto del topic ma dalla sua struttura gerarchica.

2. Un algoritmo ad apprendimento supervisionato è una tecnica di apprendimento automatico che mira ad istruire un sistema informatico. L'apprendimento non supervisionato è una tecnica di apprendimento automatico che fornisce al sistema una serie di input che esso stesso classificherà per cercare di effettuare ragionamenti e previsioni.

3. Byron, L. and Wattenberg, M. (2008) 'Stacked graphs – Geometry & aesthetics', IEEE Transactions on Visualization and Computer Graphics, 14(6), pp. 1245–1252.

4. Grafi in cui la distanza tra i nodi è il risultato di un parametro che li accomuna.

1.2 Tra ricerca e sperimentazione

Definizione dei confini

Il primo passo è stato quello di delimitare i confini della ricerca. Prima di iniziare una ricerca, è necessario fare un'operazione di circoscrizione di limiti, al fine di focalizzare l'attenzione sullo specifico settore e tema d'indagine:

> *Setting the boundaries of a system and the scale of its description could radically change the information we gathered from the observation. This choice should be done according to the purpose of the investigation*
>
> — *G.Scagnetti. et al. 2007*[5]

☞ Capitolo 2 e 3

5. Scagnetti, G. et al. (2007) 'Reshaping communication design tools. Complex systems structural features for design tools.', Proceedings of IASDR 07, (November).

Tra machine learning e data visualization

Una volta definiti i confini, non essendo reperibile in letteratura una rassegna di casi di visualizzazioni di risultati di algoritmi di *topic modeling*, il volume presenta due capitoli dedicati al contesto teorico della ricerca sul *topic modeling* che mettono in luce lo stato dell'arte relativo a questo tema. Ad una prima rassegna dedicata alla storia degli algoritmi di *topic modeling* seguirà un'analisi più specifica, connessa ai modelli visivi più ricorrenti per rappresentare questo genere di dati.

☞ Capitolo 3 e 4

Progetto Topic Tomographies

Il capitolo 5 è riservato al progetto sperimentale di partenza, *TopTom - Topic Tomographies*. Trattandosi di un progetto interdisciplinare, la cui prima parte dedicata all'estrazione e all'elaborazione dei dati è stata condotta da un team di data scientists, una sezione relativa alla struttura del dato è necessaria per comprendere la complessità del progetto e familiarizzare con un tema che verrà affrontato nell'ultimo capitolo di questo volume .

Topic Tomographies è una piattaforma digitale il cui obiettivo è quello di fornire soluzioni analitiche e visive per l'esplorazione di un *corpus* dinamico di contenuti provenienti dal web[6]. Per garantire una buona esplorazione del dato, lungo multiple scale temporali e diverse strutture gerarchiche sono state adottate metafore visive provenienti dal campo biologico e medico.

Particolare attenzione sarà dedicata all'aspetto di UI/UX[7] e ad un'analisi critica del risultato ottenuto, con l'ottica e la speranza di poter integrare, in un'altra occasione, il progetto *TopTom* .

Sperimentazione

Sia dalla ricerca bibliografica che dagli *insights* emersi durante il progetto *TopTom* è sorta la necessità di indagare a fondo il tema della rappresentazione del topic. Già l'etimologia del termine topic, dal greco τόπος, τόπος κοινός non ne nasconde la complessità: il topic, infatti, prima di essere un argomento, è un *luogo comune a più elementi.*

Il topic in sé è un concetto astratto, privo d'identità, esistente solo quando definito da parametri che ne sottolineano la struttura complessa.

Obiettivo della sperimentazione è stato identificare un parametro che definisse un aspetto importante dell'identità del topic e visualizzare un set d'esempio di elementi. A seguito di numerosi tentativi è emerso che un metodo interessante per visualizzare l'identità di un topic è fare riferimento alla struttura dei dati che lo definiscono lasciando momentaneamente in disparte il contenuto testuale.

☞ Capitolo 5

6. Nello specifico contenuti prodotti dagli utenti provenienti da Twitter, Reddit e GDelt.

7. Interfaccia ed esperienza utente.

☞ Capitolo 6

Visualization leverages the human visual system to augment human intellect.

— Mike Bostock, Visualizing Algorithms, 2013

2. Visualizzazione dati e topic modeling

La nascita del rapporto tra *data science*[8] e design dell'informazione[9] è da considerarsi recente e legata al progresso tecnologico di ambo i settori. Il sempre più rapido miglioramento degli algoritmi di *machine learning* e l'ottimizzazione dei linguaggi di programmazione front-end hanno contribuito, negli ultimi dieci anni, ad una crescita esponenziale dei casi di collaborazione tra data scientists, statistici, programmatori e designer dell'informazione. Nel capitolo saranno mostrate alcune analisi, svolte tramite l'utilizzo dei *Digital Methods*[10], con l'obiettivo di fornire un contesto d'azione alla ricerca. La sezione introduttiva sarà dedicata all'analisi dei gruppi di *seealso* di coppie di pagine Wikipedia, poi, da un focus generico tra *information design* e *data science* si passerà ad un focus più specifico tra *data visualization* e *topic modeling*. La seconda sezione, sarà dedicata alle pubblicazioni scientifiche sul tema *topic modeling* che hanno coinvolto la *data visualization*.

8. La scienza che studia le modalità di estrazione della conoscienza a partire dai dati.

9. Ampio settore del design che si occupa della rappresentazione delle informazioni.

10. *The Digital Methods Initiative (DMI) is a contribution to doing research into the natively digital.*

2.1 *Seealsology*: un primo sguardo

Molte pagine di Wikipedia annoverano la sezione *Pagine correlate*, in inglese *Seealso*.
Attraverso l'analisi della sezione *Seealso* di una o più pagine, Wikipedia offre la possibilità di indagare ad un livello superficiale come siano correlate tra loro diverse tematiche. Solitamente è preferibile affrontare l'analisi confrontando i risultati per diverse lingue, ma in questo caso non sarebbe stato particolarmente rilevante. *Topic model* è un'espressione usata esclusivamente in lingua inglese, pertanto non esistono articoli di Wikipedia tradotti nelle lingue con alfabeto latino, fatta eccezione per la pagina francese (https://fr.wikipedia.org/wiki/Topic_model), la quale però ha uno scarno numero di contenuti rispetto a quella inglese. Le uniche traduzioni esistenti sono per il persiano, coreano, cinese e russo.

Lo strumento open source *Seealsology*,

> *A simple tool that allows you to explore in a quick and dirty way the semantic area related to any Wikipedia Page.*
>
> — *Medialab SciencèsPo , sito ufficiale*[11]

consente di ricavare la rete di connessioni di seealso tra due pagine.
Il network (fig. ☞ 01) è stato generato con *Gephi*, un software sviluppato dal *Medialab SciencèsPo* di

11. http://tools.medialab.sciences-po.fr/

Parigi che consente di creare, modificare ed esportare networks. Sia per questa visualizzazione che per la successiva (fig. ☞ 02) sono stati impostati gli stessi parametri al fine di poterne confrontare struttura e contenuto. L'uso del *Force Atlas Layout*[12] unito alla *modularity*[13], ha permesso di evidenziare gruppi di pagine mettendo in luce le relazioni e la vicinanza tra di esse.

12. Layout derivato dal modello più generale force directed graph.

13. Calcolo algoritmico applicabile in Gephi che identifica clusters di nodi e permette di differenziarli cromaticamente attraverso l'interfaccia del software.

Information Design e Data Science

Osservando il primo network (fig. ☞ 01) emerge come dominante l'insieme di nodi legati all'ambito della visualizzazione dati, in cui la pagina *Data Visualization* è la più ricca di connessioni. Il gruppo attorno alla pagina *Data Science* è di dimensioni ridotte e connesso alla pagina *Information Design* solo attraverso il gruppo di nodi formatosi attorno alla pagina *Statistics*. Il cluster attorno alla pagina *Data Visualization* è connesso direttamente al mondo della *data science* dalle *pagine-ponte Data Analysis* e *Big Data*.
Le relazioni più forti sono *data visualization-data science*, *information architecture - content management - statistics*, ed infatti, il progetto *TopTom* ruota prevalentemente attorno a questi quattro settori.

Data Visualization e Topic Modeling

Emersa come pagina dominante, *Data Visualization* è stato il punto di partenza per una seconda analisi di contesto, più specifica ed inerente alla tematica della tesi: la relazione tra *topic modeling* e visualizzazione dati. Inoltre durante l'ultima edizione di *IEEE VIS*, la più importante e conosciuta conferenza al mondo riguardo visualizzazione per le scienze, visualizzazione delle informazioni e analisi visive, la pubblicazione *Progressive Learning of Topic modeling Parameters* è stata premiata come miglior paper.

LEGENDA

una pagina di Wikipedia

quantità di seelaso
della pagina

ambiti specifici
di interesse per la tesi

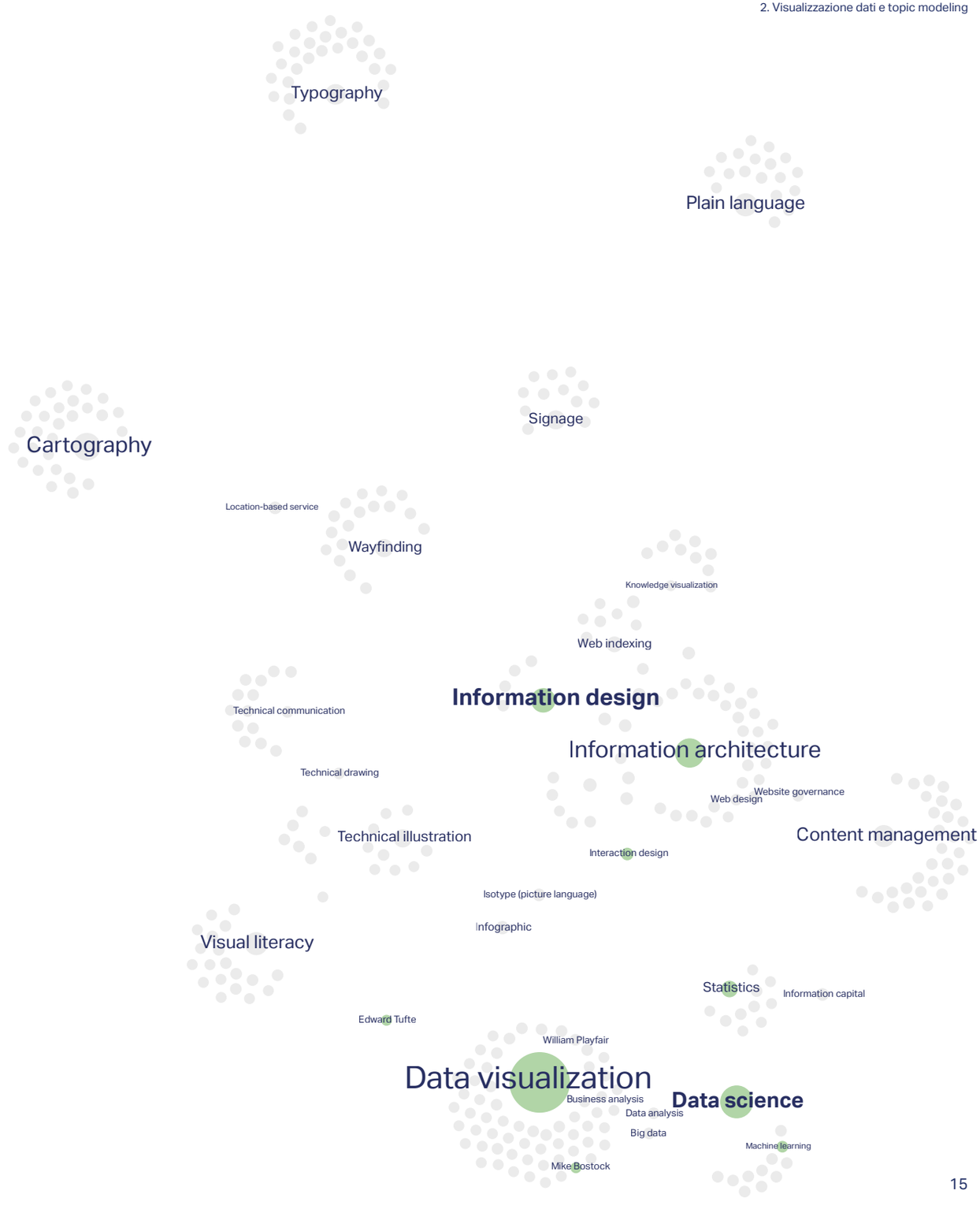


fig.01 Rete generata da dati estratti con Seealsology a partire dalle pagine di Wikipedia *Data Science* e *Information Design*. L'output è stato generato con il software Gephi, utilizzando il Force Atlas Layout. L'immagine mostra come l'abito del design dell'informazione più connesso alla *data science* sia quello della data visualization. I nodi colorati in verde rappresentano le tematiche che verranno affrontate in questo volume.

LEGENDA

una pagina di Wikipedia

quantità di seelaso
della pagina

ambiti specifici
di interesse per la tesi

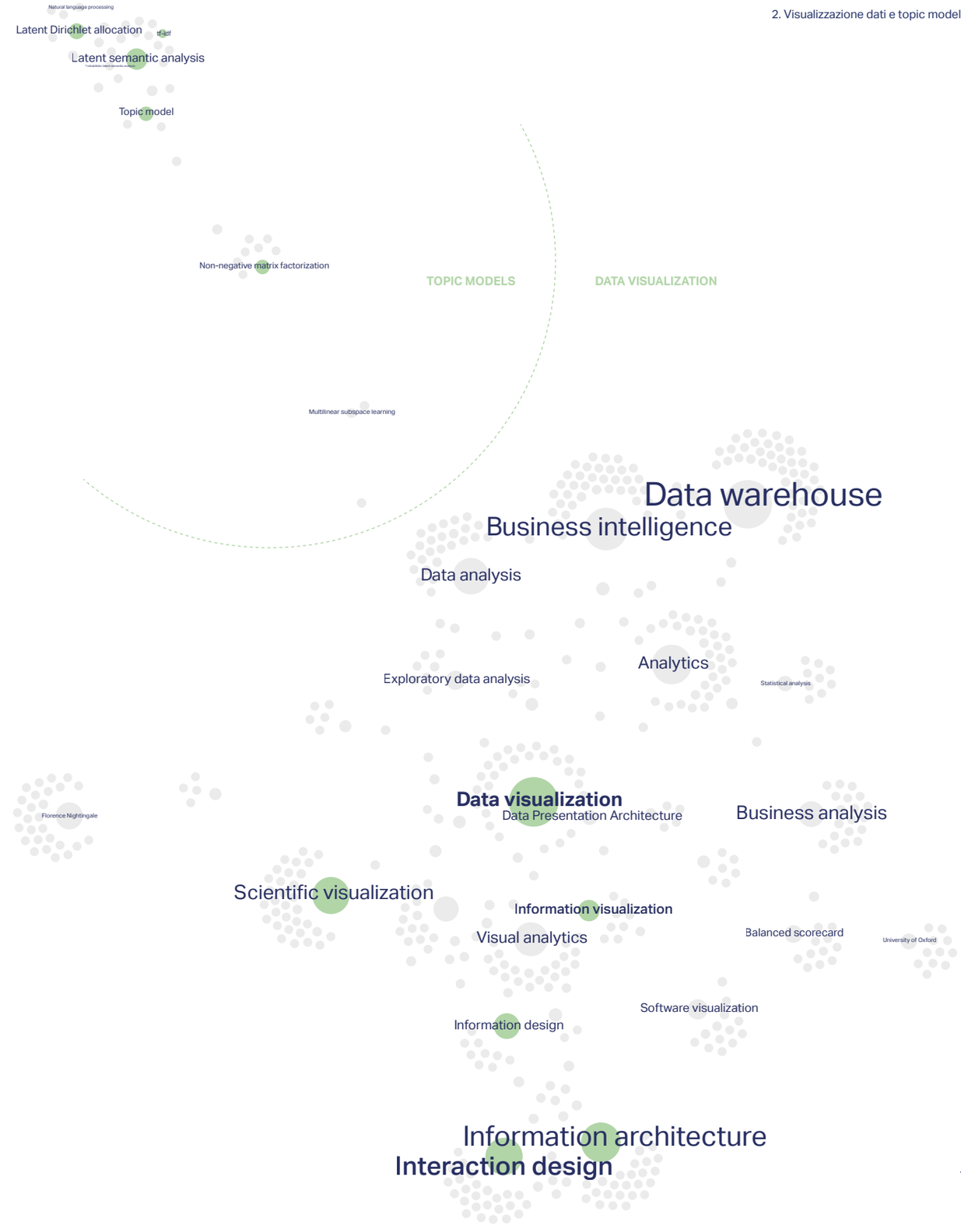


fig.02 Rete generata da dati estratti con Seealsology a partire dalle pagine di Wikipedia *Topic_Model* e *Data_Visualization*. L'output è stato generato con il software *Gephi*, utilizzando il *Force Atlas Layout*. L'immagine mostra come l'ambito del *topic modeling* e della visualizzazione dati siano distanti e condividano pochi nodi. I nodi colorati in verde rappresentano le tematiche che verranno affrontate in questo volume.

Osservando le immagini (fig. ☞ 01 e 02) come prima cosa, salta all'occhio le strutture differenti dei network. Se il network tra *Information Design* e *Data Science* (fig. ☞ 01) ha un aspetto compatto e la relazione tra i due *seeds*[14] è un insieme di pagine interconnesse tra loro, nel caso della seconda visualizzazione (fig. ☞ 02) è evidente come i due blocchi, il primo legato al *topic model* e il secondo legato alla *data visualization*, siano diametralmente opposti e connessi soltanto da poche pagine.
Il principale *nodo-ponte*[15] è la pagina riguardante la *Non-negative Matrix Factorization*, uno dei principali algoritmi necessari per l'elaborazione dati nel *topic modeling* utilizzato anche nel caso di *TopTom*.

Tuttavia l'analisi del network delle correlazioni in Wikipedia, dà solo un'immagine superficiale del legame tra i diversi ambiti.
Per comprendere meglio come si è evoluta negli anni la relazione tra *data visualization* e *topic modeling* è utile affidarsi ad altri strumenti più peculiari.

14. Il nodo(i) della(e) pagina(e) principali che originano il grafo.

15. In inglese, "bridge node" è un nodo che connette due cluster.

2.2 Scopus: la comunità scientifica

Scopus[16] è

> *the world's largest abstract and citation database of peer-reviewed research literature: scientific journals, books and conference proceedings*
>
> — *Scopus , sito ufficiale*

Fornisce statistiche su alcune delle migliori pubblicazioni scientifiche, categorizzate per ambiti e sotto-ambiti. Registrandosi al portale con una e-mail accademica è possibile accedere a statistiche, ottenere metadati delle singole pubblicazioni e scaricare i documenti disponibili in formato .pdf.
La piattaforma consente inoltre di consultare abstract e parole chiave di descrizione che sono state associate all'articolo in fase di pubblicazione.
La sezione *Analyze Search Results* offre la possibilità di consultare statistiche relative ad anno di pubblicazione, autori e centri di ricerca.

Tuttavia è necessario tenere in considerazione che:

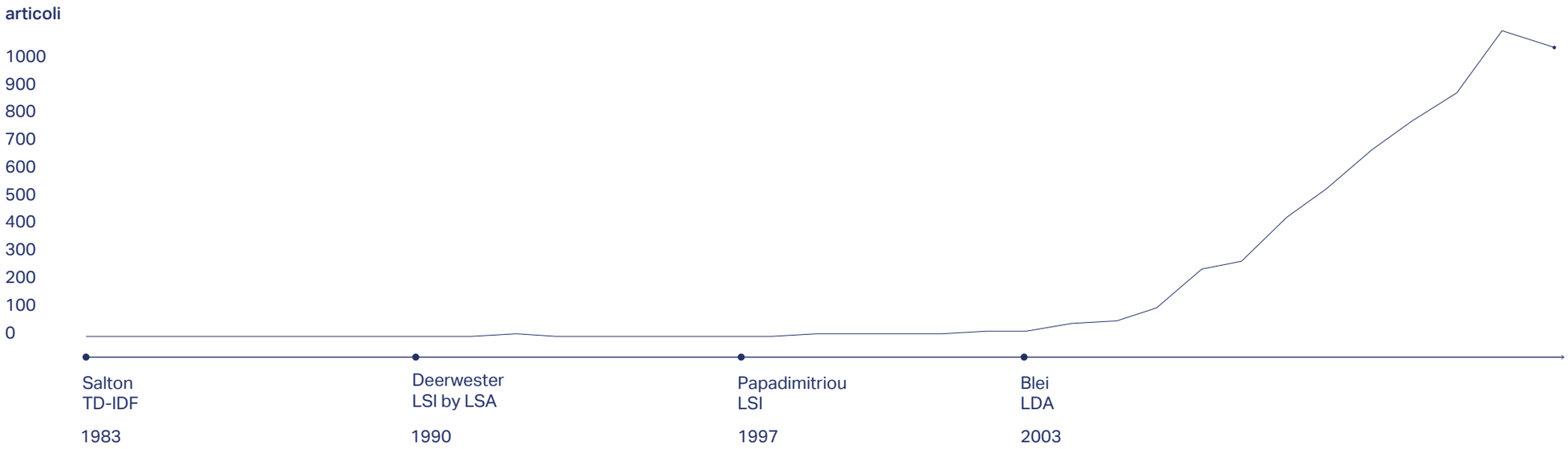
☞ Non contiene tutti i paper scritti, ma soltanto una selezione dei migliori;

☞ Per quel che riguarda le pubblicazioni antecedenti a fine anni 90', è ancora in corso un'operazione di digitalizzazione dell'archivio.

16. https://www.scopus.com/

17. https://www.elsevier.com/solutions/scopus

fig.03 Qui in alto è mostrato il numero crescente di articoli pubblicati dal 1983 al 2017 contenenti i termini “topic model” o “topic modeling” o “topic models” nell’abstract di Scopus. Sull’asse x del tempo sono segnati alcuni papers scientifici (e relativi autori) che hanno segnato la storia del *Topic modeling* (fonte : Scopus)



Quarant'anni di progresso

Nell'arco degli ultimi quarant' anni sono stati registrati un totale di 5695 documenti/papers/articoli il cui abstract, titolo o sottotitolo contiene i termini "topic model" OR "topic modeling" OR "topic models", soprattutto a partire dal 2003, anno in cui David Blei ha pubblicato il paper *Latent Dirichlet Allocation*[18]. (fig. ☞ 03)

Non essendo il risultato ancora soddisfacente, perché estremamente generico, affidando il processo di *query*[19] *design* alla ricerca di partenza effettuata su Wikipedia, è stata modificata la *query* di ricerca in "topic model" OR "topic modeling" OR "topic models" AND "data visualization".

L'archivio di *Scopus* contiene soltanto 33 documenti che soddisfano queste caratteristiche e che quindi sono specificatamente legati al campo della data visualization. I risultati aumentano se si modifica la *query* in "topic model" OR "*t*opic modeling" OR "topic models" AND "visualization" OR "data visualization": in questo caso i risultati sono 225 e i contenuti inerenti a progetti in cui è evidente il contributo della visualizzazione dati. In entrambi i casi è chiaro quanto la relazione tra *topic modeling* e data visualization sia debole ma in crescita, e quindi un ambito in cui la ricerca ha potenzialità.

Il contributo del design nella storia

La ricerca relativa al *topic modeling* è ancora legata all'aspetto scientifico statistico e la visualizzazione non viene sfruttata come strumento di ricerca e approfondimento ma solo con il l'obiettivo di visualizzare i dati in fase finale. (Una riflessione più dettagliata sull'argomento sarà oggetto del capitolo 3 e del capitolo 4).

In ambiti di ricerca complessi in cui molte discipline collaborano su singoli progetti, è chiaro come il ruolo della data visualization sia di estrema importanza.

Se nel 2005 soltanto su una pubblicazione appariva il termine *visualization*, alla fine del 2017 sono 50 i casi di articoli in cui la visualizzazione è citata in prima linea, per un totale di 225 papers nell'arco di 12 anni. Interessante crescita, che però corrisponde al solo 0,04% del totale dei papers pubblicati sul *topic modeling*. Interessante è osservare come questi articoli siano disseminati per centri di ricerca in tutto il mondo: Europa, Cina e Stati Uniti pubblicano la maggior parte di articoli relativi al tema (fig. ☞ 04), ma nemmeno la metà annoverano anche i termini *data visualization* o *visualization.* Usando la *media* la situazione cambia (fig. ☞ 05). Emerge come nei centri di ricerca cinesi si tenda raramente ad inserire *data visualization* nelle parole chiave che identificano l'articolo rispetto ai centri di ricerca americani. Caso interessante, è l'Italia che si posiziona all'ultimo posto, avendo utilizzato soltanto una volta su 105 pubblicazioni il termine data visualization nella descrizione dell'articolo.

18. Blei, D. M. et al. (2003) 'Latent Dirichlet Allocation', Journal of *Machine learning* Research, 3, pp. 993–1022. doi: 10.1162/jmlr.2003.3.4–5.993.

19. In informatica, indica il termine che l'utente sta cercando all'interno di un database.

LEGENDA

1764
1

numero di articoli scritti

articoli contenenti *topic models…*

articoli contenenti *data visualization …*

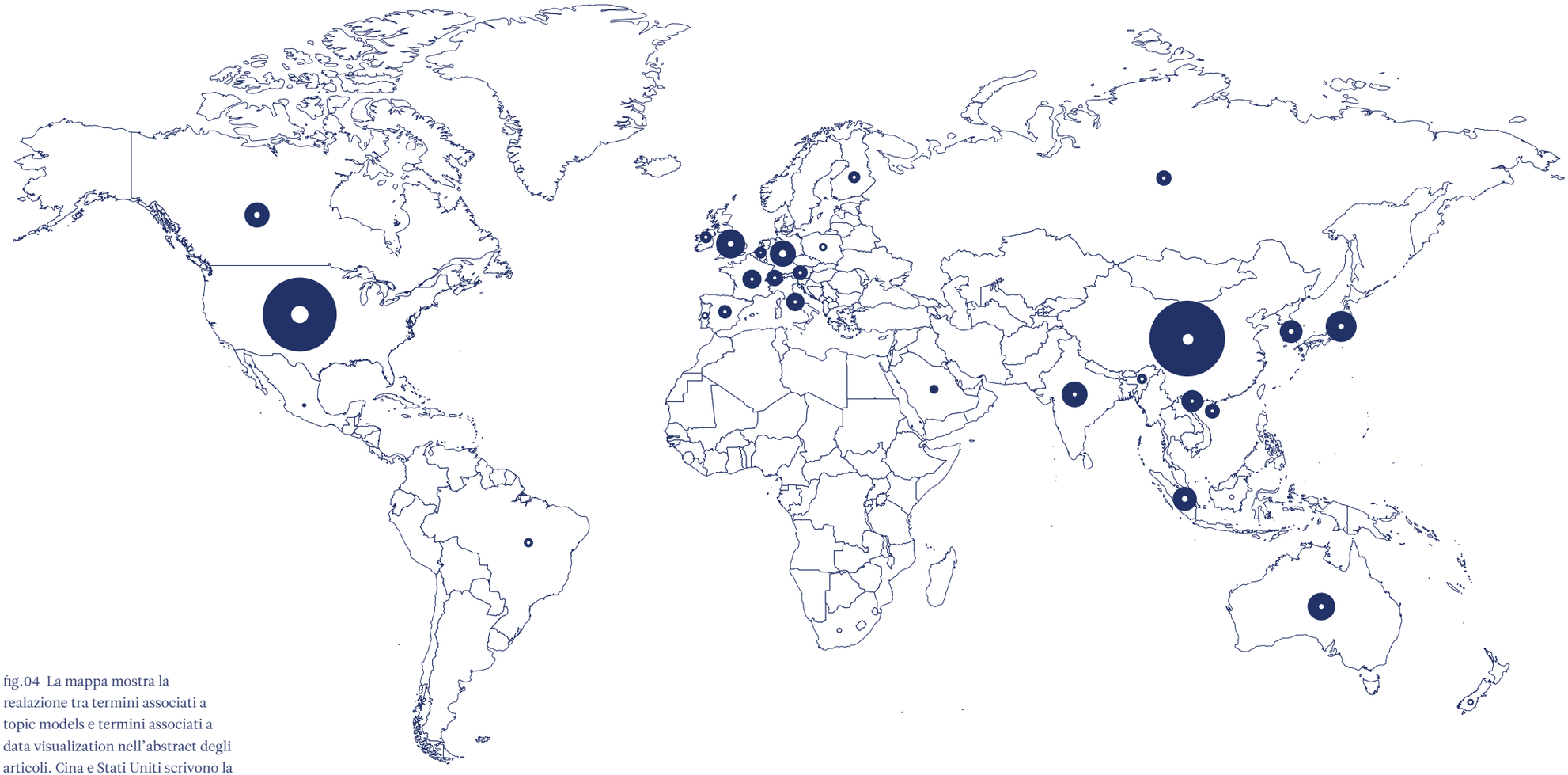


fig.04 La mappa mostra la realazione tra termini associati a topic models e termini associati a data visualization nell'abstract degli articoli. Cina e Stati Uniti scrivono la maggior parte degli articoli, non tenendo molto in considerazione i termini *data visualization* e *visualization*.

LEGENDA

Proporzione tra numero di papers dedicati esclusivamente al topic modeling e papers con data visualization

100% 0%

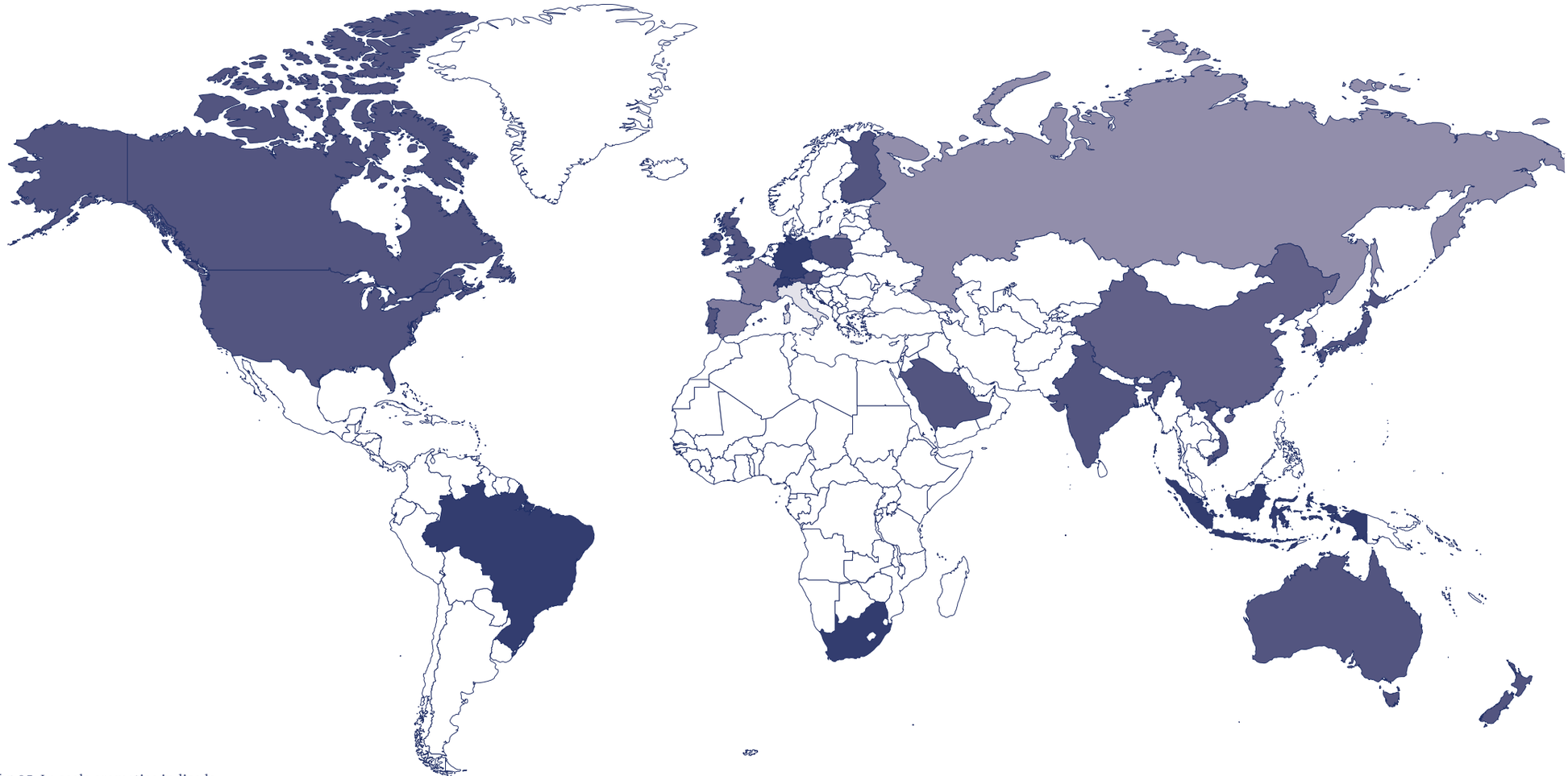


fig.05 La scala cromatica indica la proporzione di papers scientifici che trattano contemporaneamente l' argomento *topic modeling* e *visualizzazione dati*.

2.3 Verso una bibliografia consistente

L'analisi presentata in questo capitolo serve solo a mostrare un background della ricerca e del progetto sperimentale presentato in questo elaborato di tesi. Per mappare il fenomeno nella sua interezza sarebbe necessario analizzare altre piattaforme come, ad esempio, *Google Scholar*[20] o *Academia.edu*[21] oppure cataloghi specifici di articoli specifici (*IEEE Transactions on Visualization and Computer Graphics*[22]) e confrontare i risultati con quelli mostrati nelle pagine precedenti. Tuttavia, *Scopus* si è dimostrata una piattaforma da cui è stato possibile ricavare sia informazioni legate all'andamento generale del fenomeno sia una buona base di casi studio da analizzare come punto di partenza della ricerca bibliografica che sarà approfondita nel capitolo successivo.

20. Motore di ricerca accessibile liberamente che tramite parole chiave specifiche consente di individuare testi della cosiddetta letteratura accademica.

21. Sito web per ricercatori dedicato alla condivisione delle pubblicazioni scientifiche.

22. Rivista mensile di computer science e visualizzazione dati

**Good artists copy,
great artists steal.**

— Pablo Picasso, 1892

3. Luoghi sconosciuti

3.1 I Topic Models

> *What exactly is a* topic*? Formally, a* topic *is a probabilty distribution over terms*
>
> — *D. Blei, 2003*

I *topic models* sono insiemi di algoritmi di *machine learning*[23] che permettono di riconoscere i temi trattati all'interno di un ampio numero di documenti. Possono essere applicati a varie tipologie di dati non strutturati dei quali ne agevolano categorizzazione

23. Campo della computer science che da ad un computer la capacità di imparare senza essere specificatamente programmato.

e sintesi, processi altrimenti impossibili con il solo apporto fornito dalla *human categorization*[24].
L'obiettivo degli algoritmi di *topic modeling* è quello di identificare i topic attraverso la co-occorrenza di parole a partire da un *corpus* di documenti.
Gli elementi coinvolti in un modello statistico di *topic modeling* sono:

- Documenti
- Topic
- *Keywords*

Un *documento* può essere un tweet, un articolo, una pubblicazione, un libro.
Un insieme di documenti è il *corpus*. Tutti i documenti del *corpus* possono trattare argomenti, ovvero topic, più o meno comuni agli altri documenti del *corpus*.
Le parole chiave, *keywords*, presenti nei documenti definiscono i topic. L'insieme di *keywords* presenti nei *documenti* compone la cosiddetta *bag of words*.
I passaggi tipici dei più recenti algoritmi di *topic modeling* sono riassumibili in cinque o sei step principali.

Raccolta dei documenti
Una prima fase consiste nella raccolta dei documenti da analizzare: possono essere un gruppo di tweets relativi a determinate *query*, come nel caso di *Mining concurrent* topic*al activity in microblog streams* (A. Panisson et al., 2014) in cui sono stati raccolti tutti tweets relativi alle Olimpiadi di Londra del 2012, oppure papers scientifici come nel caso di *Termite : Visualization Techniques for Assessing Textual* Topic *Models* (J. Chuang et al., 2012) di che analizza 372 conference papers della IEEE InfoVis.

24. Processo di classificazione effettuato dagli esseri umani

Pulizia delle keyowords in eccesso
Se si prendessero tutte le parole senza nessun tipo di elaborazione, il *topic modeling* considererebbe anche articoli, preposizioni e avverbi. Uno step fondamentale è infatti quello di definire un dizionario privato dalle *stop words*[25].

Valutazione relazione topic/keywords/documenti
A questo punto si procede con la valutazione delle singole parole in relazione sia al documento di appartenenza che al *corpus* di documenti per intero.

Definizione dei topic
La definizione dei topic è il passaggio successivo in cui si definiscono le parole che caratterizzano i topic e i documenti a cui ciascun topic si riferisce. Un documento può trattare più topic in diverse proporzioni, così come una *keyword* può appartenere a diversi documenti e diversi topic in diverse proporzioni.

Topic modeling dinamico
In alcuni casi si può considerare anche la variabile temporale, per vedere l'evoluzione dei topic lungo il tempo.

Topic modeling gerarchico
Inoltre, nell'arco di una giornata, di un'ora o di una settimana è possibile identificare diversi livelli di dettaglio del topic sulla base dello stesso *corpus*.
In questo modo si potrà dire che in una giornata si è parlato genericamente di *oppiacei* oppure fare riferimento a temi più specifici come, ad esempio, *la somministrazione dell'epinefrina*[26].

25. Un avverbio, congiunzione, preposizione, pronome, articolo...

26. Questi esempi specifici fsnno riferimento al *corpus* d'esempio utilizzato nel progetto TopicTomographies

3.2 Dalla mente umana al machine learning

Come ogni algoritmo di *machine learning*, così anche gli algoritmi di *topic modeling* sono soggetti a innumerevoli modifiche e migliorie anno dopo anno. Tuttavia, prima di parlare di *topic models* come modello statistico probabilistico è necessario andare indietro nella storia e definire quali siano stati gli eventi e le teorie che hanno contribuito alla definizione ed evoluzione di questo modello.
Infatti ciò che oggi è comunemente noto come *topic modeling* pone radici nella pratica dello *human classification*. Fino a poco più di vent'anni fa ottenere informazioni relative a molti documenti era il risultato di un'operazione di lettura e classificazione da parte di esseri umani, infatti l'abilità di categorizzazione degli esseri umani è considerata la base dell'apprendimento e della conoscenza nelle ricerche sulla *human classification*.

3.2.1 Le prime contaminazioni

Il primo esperimento analogico che si può ricondurre al *topic modeling* è un esperimento di Charles Spearman del 1909.
Spearman fu uno psicologo inglese conosciuto prevalentemente per studi statistici sull'intelligenza e l'apprendimento.

Durante un esperimento del 1909 chiese ad alcuni volontari di associare una collezione di dieci documenti a una lista di argomenti da lui stesso stilata, assegnando un valore da 1-10 sulla base di quanto ciascun documento trattasse le singole tematiche in lista.
Dato il background di Spearman il fine specifico dell'esperimento era comprendere in che modo diversi individui categorizzano contenuti, non porre le basi per gli studi sulla *human classification*. Tuttavia, è estremamente interessante notare come la struttura di rappresentazione usata da Spearman sia in realtà una matrice mxn al cui incrocio tra argomenti e documenti c'è un valore che descrive quanto ogni singolo documento tratti quell'argomento, e questa struttura è alla base dei più recenti algoritmi di *topic modeling*.

3.2.2 Parole, parole, parole

Strumenti necessari per comprendere i principi base degli avanzati strumenti di *topic modeling*.
Era tuttavia necessario trovare un metodo matematico-statistico per quantificare la frequenza dei termini nei documenti, definendo degli indici.
Nel 1975 Gerard Salton sostiene che

> *the basic assumption in any text data mining method is that a document can be represented as a vector of terms in the vector space model (VSM)*[27]
>
> — *G. Salton, 1975*

dove l'importanza di un termine è denotata dalla frequenza con cui appare in un documento, rappresentata nello spazio.
Partendo da un set di esempio di tre documenti

27. Salton, G., Wong, A. & Yang, C. S. A vector space model for automatic indexing. Commun. ACM 18, 613–620 (1975).

☞ D1: La casa è sul fiume

☞ D2: La casa nel bosco è un rudere

☞ D3: Nel bosco c'è un rudere

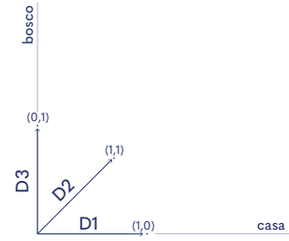
Per comprendere la rilevanza all'interno del *corpus* dei termini *casa* e *bosco*, ovvero le *query* di ricerca, è necessario organizzare i documenti secondo una matrice mxn (fig. ☞ 06 e 07) in cui le righe sono i documenti e le colonne sono le *query*.
Quasi contemporaneamente Salton teorizza anche il *Tf-idf scheme*, un calcolo in grado di pesare l'importanza dei termini all'interno dei documenti assegnando loro un valore numerico. (fig. ☞ 08)
Nel *VSM* infatti non c'erano riferimenti alle dimensioni del documento ne del *corpus*, nonostante sia importante per capire quanto un termine sia rilevante all'interno di un *corpus* di documenti.
Per esempio se su 100 parole totali di un documento, la parola casa appare 5 volte, questa avrà rilevanza 5/100 = 0,05. Se in tutto il *corpus*, composto da 1000 documenti, la parola casa appare in 10 documenti avrà rilevanza di log(1000/10) = log 100= 2 e complessivamente il termine casa avrà una rilevanza pesata su tutto il *corpus* di 0,05 x 3 = 0,1.

fig.06 La matrice mxn alla base del Vector Space Model. Ogni documento è valutato in base alla presenza di alcune *query* al suo interno.
fig.07 I valori di ogni documento sono posizionati nello spazio *x,y,z*
fig.08 Uno schema del term frequency (sinistra) e documents frequency (destra) alla base dello schema tf-idf.

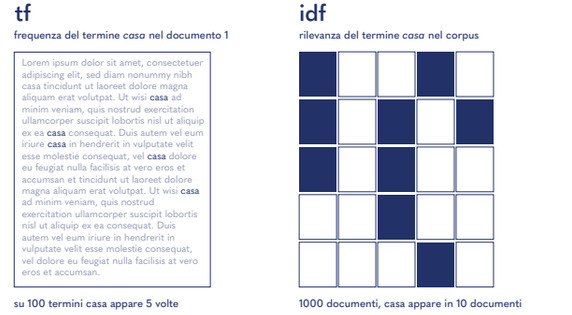
	casa	bosco
La casa è sul fiume	1	0
La casa nel bosco è un rudere	1	1
Il rudere è nel bosco	0	1

06

	Q1 (x)	Q2 (y)
D1	1	0
D2	1	1
D3	0	1



07



08

2003 · LDA
Latent Dirichlet Allocation
Blei

1999 · pLSI
Probabilistic Latent Semantic Analysis
Hofmann

1990 · LSI
Indexing by latent semantic analysis
Deerwester, Furnas et al.

1983 · tf-idf
Extended Boolean information retrieval
Salton et al.

1975 · VSM
A vector space model for automatic indexing
Salton et al.

1909
Esperimento di Charles Spearman

fig.09 Le principali pubblicazioni legate al *topic modeling*, nello specifico quelle citate da David Blei nella pubblicazione del 2003.

3.3.3 Topic latenti

Indexing by latent semantic analysis (LSA)

In *Indexing by latent semantic analysis* (S. Deerwester et.al, 1990) gli autori sostengono che è possibile, attraverso il loro nuovo modello derivato dal preesistente *tf-idf* (vedi sezione precedente), considerare alcuni aspetti della linguistica di base come i casi di *sinonimia*[28] e *polisemia*[29].
Per questo motivo viene introdotto il concetto di *latenza* per la prima volta sottolineando che *esiste una certa struttura semantica latente, parzialmente nascosta dalla casualità della scelta delle parole.*
Il modello, chiamato sia *LSI - Latent Semantic Indexing* che *LSA - Latent semantic Analysis* si basa su una matrice *terms x docs* in cui il valore all'incrocio corrisponde al tf-idf del termine.
A differenza del modello base di Salton del 1983, il modello di Deerwester analizza i valori dei termini appartenenti alla matrice *terms x docs* nel Vector Space Model, in modo da ottenere una rappresentazione più specifica dei lemmi, tenendo in considerazione la similarità tra termini, in altre parole, sinonimia e polisemia.
Tuttavia oggigiorno, gli algoritmi di *topic modeling* si basano sul principio della probabilità, e anche il più avanzato modello di Deerwester non teneva ancora in considerazione questo fattore.

Probabilistic Latent Semantic Analysis (pLSA)

Nel 1999 è Thomas Hoffman, dall'International Computer Science Institute, Berkeley, CA ad affrontare per primo il tema della probabilità nel *topic modeling* e, soprattutto, è il primo ad introdurre il concetto di topic. (T. Hofmann, 1999)

The main challenge a machine learning system has

28. Identità quasi sostanziale di significato tra due o più parole o espressioni,

29. La coesistenza, in uno stessa parola, di significati diversi, ad esempio vite, sia come oggetto metallico che pianta.

> *to address roots in the distinction between the lexical level of "what actually has been said or written" and the semantic level of "what was intended" or "what was referred to" in a text or an utterance.*
>
> — *T. Hoffman, 1990*[30]

Non è possibile definire in maniera arbitraria quali parole appartengano ad un determinato argomento, o topic, proprio perché la stessa parola può appartenere a contesti del tutto differenti. Infatti, Hoffman continua:

> *[...] different words may have a similar meaning, they may at least in certain contexts denote the same concept or – in a weaker sense – refer to the same* topic.
>
> — *T. Hoffman, 1990*

Nel modello teorico di Hoffman l'obiettivo principale è quello di descrivere la probabilità che un topic, definito da probabili termini, appaia in un documento.

Latent Dirichlet Allocation (LDA)

Dal 2003 a oggi il modello base di David Blei (e i successivi sviluppi dello stesso, che non verranno elencati per questioni di rilevanza ai fini della ricerca, ma che hanno contribuito al miglioramento del modello) chiamato *Latent Dirichlet Allocatio*n è uno dei più usati, e la relativa pubblicazione *Latent Dirichlet Allocation*[31] è una delle più citate su Google Scholar con uno score di 22000 punti. Secondo questo metodo le parole di ogni documento sono realizzazioni di composizioni percentuali di argomenti (topic). In particolare, ogni argomento si concretizza in una distribuzione di probabilità su un *vocabolario prefissato*; inoltre l'argomento (topic) è compatibile con ogni documento del *corpus* ma la sua frequenza all'interno del documento varia statisticamente tra i documenti. Infatti, ciò che Blei ripete molto spesso è che in una collezione di documenti esista una struttura sottesa, latente e *hidden* ad ogni topic, e partendo dalle assunzioni che:

- Documenti con topic simili useranno simili gruppi di parole;

- Ogni documento contiene un certo numero di topic e ha una distribuzione associata di parole.

Per cui, la domanda corretta da porsi prima di identificare la struttura di topic relativa ad un *corpus* di documenti è: qual è la struttura latente di topic che, probabilmente, ha generato il corpus di documenti? Risulta quindi necessario fare un'*operazione inferenziale*[32] definendo *a priori* i criteri di generazione probabilistica di un documento e non, come sarebbe più intuitivo pensare, del topic, ovvero, nell'ordine:

- Definire a priori il numero di parole nel singolo documento

- Definire a priori una *topic mixture* per un documento da un set predefinito di topic.

Tale operazione inferenziale è necessaria per consentire all'algoritmo di apprendere come determinati topic possano distribuirsi in un *corpus* esteso di documenti.
In *Probabilistic* Topic *Models* (D. Blei et al., 2010) Blei definisce essenziale il contributo del design per visualizzare informazioni legate ai risultati delle analisi di *topic modeling*:

30. Hofmann, T. (1999) 'Probabilistic Latent Semantic Analysis', Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence (UAI-99), pp. 289–296.

fig.10 David M. Blei

31. Blei, D. M. et al. (2003) 'Latent Dirichlet Allocation', Journal of *Machine learning* Research, 3, pp. 993–1022. doi: 10.1162/jmlr.2003.3.4-5.993.

32. Dal latino *inferre*, letteralmente significa portare dentro ed è il processo attraverso il quale da una proposizione assunta come vera si passa a una seconda proposizione la cui verità è derivata dal contenuto della prima.

How can we best display these - connessioni tra topic, *parole e documenti (nda) - connections? What is an effective interface to the whole corpus and its inferred* topic *structure? These are user interface questions, and they are essential to topic modeling*

— *D. Blei, 2010*[33]

33. Blei, D., Carin, L. & Dunson, D. Probabilistic topic models. IEEE Signal Process. Mag. 27, 55–65 (2010).

Topic modeling dinamico

Una feature importante è stata aggiunta nel 2006, quando Blei ha teorizzato il *topic modeling dinamico* (Blei et al., 2006) per identificare l'evoluzione di topic nel tempo e il loro comportamento.

Topic modeling gerarchico

Alla fine del primo decennio degli anni duemila, viene introdotto il concetto di gerarchia di topic (Blei et al., 2007). Come già accennato a pag 45, nell'arco di una giornata si possono identificare topic generici originati dalla clusterizzazione di topic più dettagliati.

3.3 Unicuique suum

Nelle ultime sezioni sono stati elencati i modelli che hanno segnato la storia degli algoritmi di *topic modeling*.
Nell'arco degli ultimi quarant'anni le analisi di contenuti di grandi quantità di documenti sono passate dall'essere eseguite manualmente da umani ad essere frutto di calcoli algoritmici sempre più performanti. Calcoli algoritmici sempre più performanti hanno richiesto, nel tempo, una sempre più accurata ricerca di modelli visivi utili a visualizzarne i risultati.

Rubare idee da una persona è plagio, rubarle da molte è ricerca.

— Arthur Bloch, 1988

4. Analisi della bibliografia di settore

4.1 Una tassonomia dei casi studio

I casi studio presi in analisi nel seguente capitolo sono progetti che affrontano il tema della visualizzazione del *topic modeling* (o forme antecedenti di analisi testuale) nell'ambito della ricerca scientifica.
Dai primi anni 2000 in poi grazie alle teorie di Blei, al progresso tecnologico e all'interesse per le discussioni su piattaforme web, la ricerca scientifica ha investito sempre di più in questo settore.

*Text Visualization Browser (*Kucher e Kerren, 2015*)* raccoglie tutte le pubblicazioni di IEEE più importanti riguardo alla visualizzazione di testi e permette di esplorare l'archivio filtrando i risultati secondo una tassonomia e il tipo di dati. Sempre aggiornata, la piattaforma vanta una funzionale organizzazione dei contenuti, tuttavia la tassonomia utilizzata è molto generica perché applicabile a tutte le visualizzazioni di testo. Poiché una tassonomia per le visualizzazioni

dei risultati di *topic modeling* non è ancora stata redatta, scopo di questo capitolo è quello di cercare di organizzare i progetti contenuti nella bibliografia secondo quattro caratteristiche tipiche dei risultati di algoritmi di *topic modeling*.

Cronologia, *contenuti*, *gerarchia* e *relazioni* sono i quattro aspetti che caratterizzano lo stato dell'arte dei progetti legati alla visualizzazione dati del *topic modeling* e ne definiscono l'efficacia.
Queste quattro ipotetiche categorie sono tutte sottocategorie del principio di macro e micro letture di Edward Tufte. In *Envisioning Information* (E. R. Tufte, 1990), afferma che mostrare la complessità dei dati è un arduo lavoro, ma consente all'utente di esplorare nel miglior modo possibile i contenuti, creando narrazioni personalizzate.

> *[...] So much for the conventional, facile, and false equation: simpleness of data and design = clarity of reading. Simpleness is another aesthetic preference, not an information display strategy, not a guide to clarity. What we seek instead is a rich texture of data, a comparative context, an understanding of complexity revealed with an economy of means.*
>
> — *E. R. Tufte, 1990*[34]

L'innovazione tecnologica in ambito di programmazione front-end combinata con le parallele migliorie in ambito UI/UX ha aperto molte porte all'esplorazione delle potenzialità di visualizzazioni dati con macro/micro letture che venticinque anni fa non si potevano neanche immaginare.
Tutti gli aspetti dei dati estratti dal *topic modeling* possono essere organizzate seguendo un'idea di lettura macroscopica e microscopica.

fig.11 E. R.Tufte

34. E. R. Tufte, Envisioning Information, 1990.

4.1.1 Variabile temporale

Cronologia
La variabile della temporalità è comparsa nella letteratura del *topic modeling* soltanto con i *Dynamic Topic Models* (Blei et al., 2006). Senza l'aspetto dinamico della temporalità la definizione e la visualizzazione dei topic risulta approssimativa, generica, frammentata e incompleta.
Infatti, i casi di visualizzazione in cui non viene considerata la temporalità sono o antecedenti al 2006 come nel caso di *Interpretation and Trust* (J. Chuang et al., 2005) oppure orientate ad una specifica analisi come *Termite* (J.Chuang et al., 2012), il cui obiettivo è trovare topic latenti nella co-occorrenza di parole in un ampio *corpus* di papers scientifici e tesi di dottorato dal 1995 al 2010 senza considerare la variabile del tempo.
Per poter parlare dei modelli visivi usati nel *topic modeling* dinamico è tuttavia d'obbligo fare un approfondimento su *Themeriver* (S. Havre et al., 1999), che sfrutta la variabile temporale per mostrare l'evoluzione di argomenti nel tempo. Non si tratta ancora di applicazione del design a risultati di algoritmi di *topic modeling* poiché il set di dati rappresentato è estratto con un tool di analisi testuale (*SPIRE*, versione del 1999), tuttavia è il primo esempio di *streamgraph ante-litteram* che la letteratura propone.
L'idea del modello visivo usato nasce dalla metafora del fiume che fluttua nel tempo e cambia le dimensioni dei singoli flussi in relazione all'importanza delle tematiche. (fig. ☞ 12)

> *The "river" flows through time, changing width to depict changes in the thematic strength of documents temporally collocated.*
>
> — *S. Havre, 1999*[35]

35. Havre, S., Hetzler, E. & Nowell, L. ThemeRiver: In Search of Trends, Patterns, and Relationships. InfoVis 99 4 (1999). doi:10.1109/INFVIS.2000.885098

Pur trattandosi di un gruppo di lavoro composto prevalentemente da sociologi e informatici è interessante notare come la ricerca di una *metafora* abbia aiutato a spiegare il modello e, parte estremamente importante, abbia contribuito alla comprensione del contenuto da parte dei due utenti a cui è stato sottoposto il tool come test

> *Participants found ThemeRiver easy to understand, giving it an average rating of 2.5 on a scale of –3 to +3.*
>
> — *S. Havre, 1999*

La teorizzazione del modello visivo sarà oggetto di una pubblicazione del 2008 (L. Byron e M. Wattenberg, 2008) e da quel momento diventerà il modello più usato in casi di visualizzazione di *topic modeling dinamico*, poiché ogni flusso mostra l'interesse relativo ad uno specifico topic e il modello visivo consente contemporaneamente lettura macroscopica e microscopica a livello temporale.

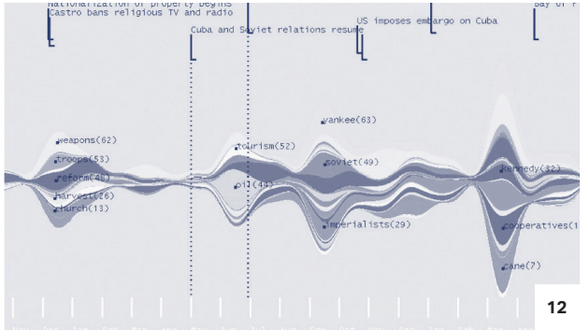
> *The main idea behind a stacked graph follows Tufte's macro / micro principle : the twin goals are to show many individual time series, while also conveying their sum.*
>
> — *L. Byron e M. Wattenberg, 2008*[36]

Soltanto nel 2010 con *Tiara: a visual exploratory text analytic system* (F.Wei et al., 2010) , con una bibliografia composta al 70% da articoli di Martin Wattenberg e Fernanda Viegas e basato su dati estratti con LDA, propone un modello visivo a *streamgraph* con una serie di interazioni che permettono di esplorare i dati.
Una caratteristica interessante ma discutibile è la chiave di ricerca tramite etichette/parole chiave che definiscono il topic nel suo totale sviluppo temporale.

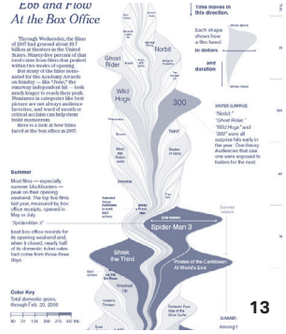
36. Byron, L. & Wattenberg, M. Stacked graphs – Geometry & aesthetics. IEEE Trans. Vis. Comput. Graph. 14, 1245–1252 (2008).

Il rischio è la generalizzazione eccessiva del contenuto del topic, colmata con il posizionamento delle etichette delle *keywords* direttamente nei singoli flussi ma visivamente in sovraccarico.
Andrè Panisson (A. Panisson et al., 2014) sperimenta lo zoom temporale tramite l'interazione sull'oggetto timeline stesso. (fig. ☞ 14)

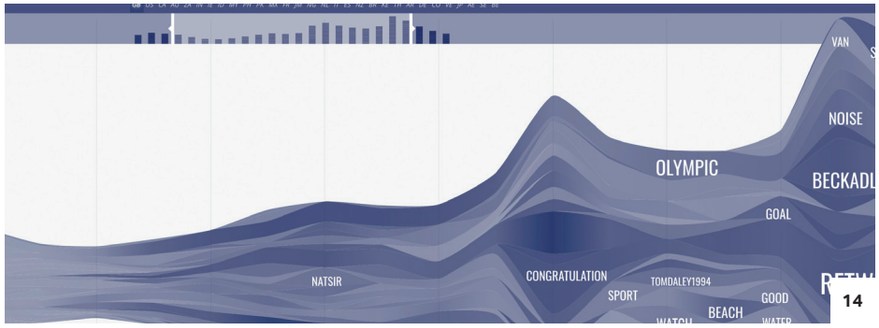
I limiti dell'aspetto cronologico, a giudicare dalla bibliografia analizzata, sono per lo più legati alla *gestione della timeline* che può essere un elemento della visualizzazione con cui interagire (click, zoom), e alla grammatica con cui si etichetta ogni singolo topic in riferimento ad un intervallo orario.
Solitamente vengono usate etichette posizionate all'interno dei flussi sebbene possano sembrare limitanti per la definizione di un topic che in realtà tratta molti
altri argomenti.



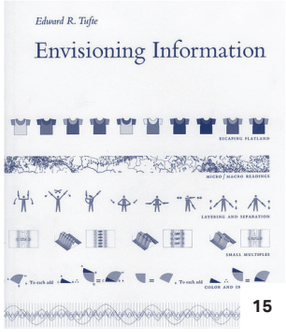
12



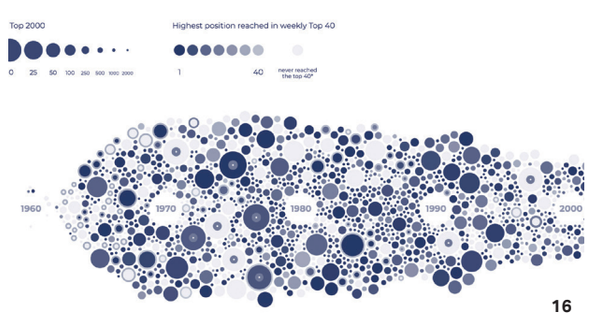
13



14



15



16

fig.12 S.Havre, *Themeriver*, *1999*
fig.13 L. Byron e M. Wattenberg, *Ebb and Flow at the Box Office*, *2008*
fig.14 A. Panisson et al., *Mining concurrent* topic*al activity in microblog streams*, 2014
fig.15 E. Tufte, EnvIsioning Information, 1990
fig.16 N. Brehmer, *Top 20 Love Songs from '80 to '0*0, 2016 (*Beeswarm* graph).

4.1.2 Struttura gerarchica

Gerarchia

Quando si parla di gerarchia negli algoritmi di *topic modeling* si fa riferimento ad un tipo di clusterizzazione gerarchica dei topic, associati per similarità secondo una struttura gerarchica ad albero che comunemente viene visualizzata attraverso un dendrogramma[37]. Gli algoritmi più conosciuti ed usati sono il *RoseTree* e la *UPGMA clusterization.*

> *The hierarchical clustering produces a tree that can be cut at a given depth to yield a clustering at a chosen level of detail.*
>
> — *W. Dou et al., 2013*[38]

A partire dal 2007 è stato possibile aggiungere questa funzionalità agli algoritmi di *topic modeling*. *Hierarchical Topics (*W.Dou et al., 2013) attraverso una piattaforma di analisi di *topic modeling* con possibilità di annotazione[39], sperimenta questo genere di algoritmo splittando l'interfaccia in tanti riquadri quanti sono i livelli di profondità gerarchica che si vogliono osservare e associando ad ogni stremagraph il corrispondente dendrogramma.

I progettisti di *#FluxFlow* (Zhao et al., 2014) (fig. ☞ 17 e 18) sperimentano un approccio molto simile a quello di *Hierarchical Topics*, usando un *beeswarm graph*[40] invece che uno *streamgraph* ed offrendo un'interfaccia più elaborata che permette di disporre verticalmente diversi livelli di aggregazione gerarchica. Come *Hierarchical Topics*, però, anche *#FluxFlow* è stato progettato come strumento di analisi e annotazione piuttosto che di monitoraggio e supervisione.

Dei ricercatori del centro Microsoft Research, in *How Hierarchical Topics Evolve in Large Text Corpora* (W. Cui et al., 2014) (fig. ☞ 19 e 20) sostengono che

> *Users can get an overview of the Hierarchical Topics evolution patterns by examining the shape changes of the color stripes and the dark regions in the bars.*
>
> — *W. Cui et al., 2014*

nonostante considerino comunque difficile ed ardua l'esplorazioni di topic in evoluzione su una struttura gerarchica.

> *However, even with coherent* topic *trees over time, the exploration and consumption of evolving* topic *in the context of hierarchies remain difficult*
>
> — *W. Cui et al., 2014*[41]

Poiché si tratta di una delle più recenti funzionalità aggiunte alla conoscenza sugli algoritmi di *topic modeling*, la letteratura su *topic modeling gerarchico e dinamico* è ancora piuttosto scarna ma, dalla bibliografia analizzata, è possibile comprendere come uno dei maggiori limiti delle visualizzazioni di *hierarchical topic modeling* sia la difficoltà di mostrare contemporaneamente diversi livelli di dettaglio dello stesso topic senza appesantire l'interfaccia.

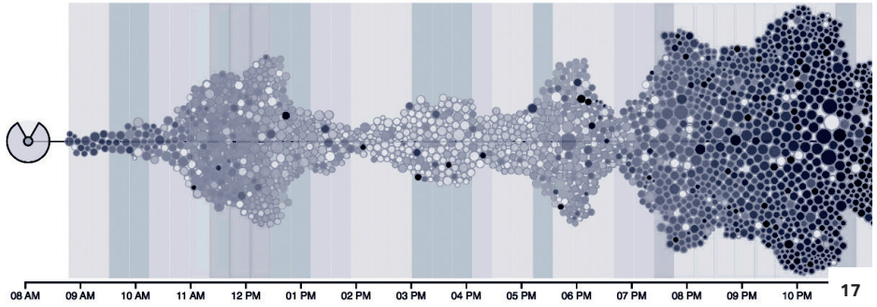
37. Modello utile per visualizzare relazioni gerarchiche o classificazione, Un esempio comune è l'albero genealogico di una famiglia.

38. Dou, W., Yu, L., Wang, X., Ma, Z. & Ribarsky, W. HierarchicalTopic: Visually exploring large text collections using topic hierarchies. IEEE Trans. Vis. Comput. Graph. 19, 2002–2011 (2013).

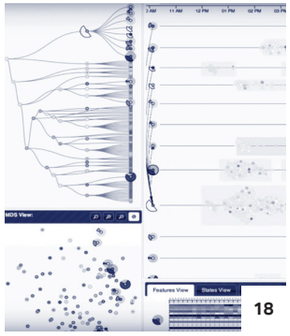
39. Alcuni tools di visual analytics consentono di annotare e modificare l'anallsi fatta dagli algoritmi

40. Modello che visualizza i dati come punti su un asse e che si espandono lungo l'altro asse per mostrare volume o quantità.

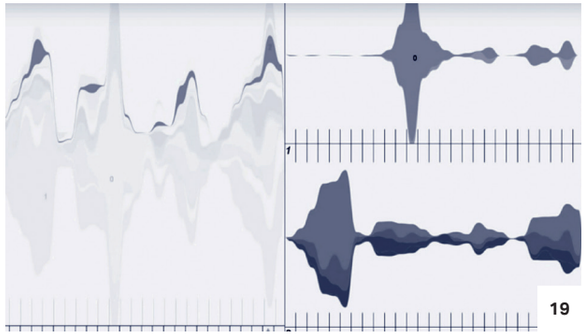
41. Cui, W., Liu, S., Wu, Z. & Wei, H. How *Hierarchical Topics* Evolve in Large Text Corpora. 20, 2281–2290 (2014).



17



18



19

20

fig.17 J. Zhao, *FluxFlow*, 2014. Vista generale del *beeswarm* dei tweets.
fig.18 Dettaglio dell'interfaccia multi-finestra di FluxFlow.
fig.19 D. Wou, *Hierarchical Topics*, 2013. Dettaglio delle viste multiple per osservare lo *streamgraph* a diversi livelli della struttura gerarchica.
fig.20 Dettaglio del dendrogramma annotabile di *Hierarchical Topics*.

4.1.3 Specificità dei contenuti

Contenuti
Assieme a gerarchia e temporalità, anche i contenuti possono essere visti a livello macro o micro. Le parole, aggregate in gruppi, compongono topic ed ogni topic può caratterizzare diversi documenti. I diversi livelli di lettura dei contenuti variano tra *keyword* / topic e *documenti*.

La letteratura annovera progetti con interfacce che prediligono la visualizzazione del numero più alto possibile di parole, come nel caso del già citato *Understanding text corpora with multiple facets* (L. Shi et al., 2010), altre visualizzazioni invece pongono il topic al primo posto come nel caso di *EvoRiver: Visual Analysis of* Topic *Coopetition on Social Media* (G. Sun et al., 2014) (fig. ☞ 22). Il focus del progetto può essere quindi più orientato sull'evoluzione di specifici pattern o sulla configurazione interna di un topic.

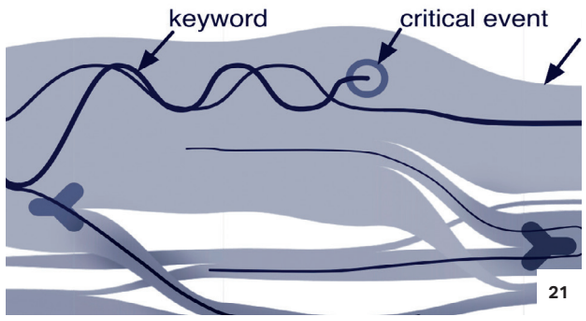
La visualizzazione dei contenuti è legata alla visualizzazione delle relazioni tra contenuti (sez. ☞ 4.1.4). Tuttavia è evidente come esista una tendenza generale ad identificare i singoli topic attraverso un'etichetta che lo descriva in termini generici (P. Xu et al., 2013, W. Cui et al., 2014) o sfruttando la *keyword* più rilevante del topic in analisi (F. Wei et al. 2010, A. Panisson et al., 2014) (fig. ☞ 24 e 25).
TextFlow: Towards Better Understanding of Evolving Topic *on Twitter* (W. Cui et al., 2011) (fig. ☞ 21), una delle pubblicazioni più citate dell'ambito specifico, oltre a considerare le *keywords* che caratterizzano un topic ne considera anche il comportamento dettagliato identificando i momenti specifici in cui un topic nasce, muore, si unisce o si divide relativamente a singole parole. Il tipo di comportamento è sottolineato dall'aspetto del *sankey diagram*[42] che varia il colore in momenti di perturbazione, fondendosi con la diramazione successiva o precedente in base al tipo di evento.
In questo caso il colore è usato per visualizzare i comportamenti, che a livello macro non è visibile ma a livello micro è identificato chiaramente. Purtroppo questo uso del colore non sfrutta appieno la potenzialità della variabile, che potrebbe essere sfruttata in altri modi. Infatti Byron e Wattenberg in *Stacked graphs - Geometry & aesthetics* sostengono che:

> *Color-coding is used to make the structure of the hierarchy vIsible. To ensure sufficient visual separation of subcategories, itwas necessary at each level to use the full range of hues. Thus, when a user changes levels in the hierarchies, the color coding must change as well.*
>
> — *L. Byron e M. Wattenberg, 2008*

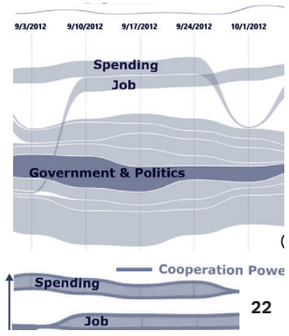
Sfruttando il colore come variabile di contenuto e allo stesso tempo di struttura gerarchica, come sarà poi sostenuto e testato in *Roseriver* (W. Cui et al., 2014) sei anni più tardi.

La configurazione dell'accesso ai documenti è di entità variabile: la maggior parte delle interfacce consente un focus sul documento sdoppiando lo schermo (S. Rönnqvist et al., 2014), (F. Wei et al., 2010) (fig. ☞ 24) (J. Zhao et al., 2014) (fig. ☞ 23).

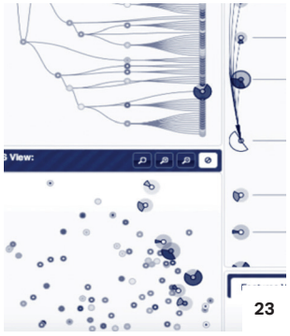
42. Il diagramma di Sankey è un particolare tipo di diagramma di flusso in cui l'ampiezza delle frecce è disegnata in maniera proporzionale alla quantità di flusso.



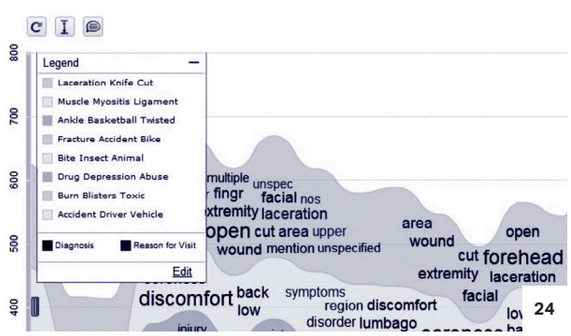
21



22



23



24

25

fig.21 W. Cui et al., *TextFlow*, 2011. Dettaglio del *sankey diagram* che si biforca in più rami. Sfumature di colore sono usate per indicare il cambiamento dei rami.
fig.22 G. Sun et al., *Evoriver*, 2014. Dettaglio della vista principale con flussi di colore diverso per ogni gruppo di topic identificato
fig.23 J. Zhao et al., *Fluxflow*, 2014. Dettaglio dell'interfaccia con pannelli multipli.
fig.24 F. Wei et al. *TIARA*, 2010. Dettaglio dello *streamgraph* con le keyword in evidenza e del pannello che raggruppa i diversi topic.
fig.25 A. Panisson et al, *Mining concurrent topical activity in microblog streams, 2014. Streamgraph* con etichetta della keyword. (http://www.datainterfaces.org/projects/emoto/)

4.1.4 Le relazioni tra elementi

Relazioni

> *Narrative meaning can be constructed from visual properties of network graphs such as topology, density of connections, absence of connections, size, position and colour of nodes.*
>
> — *Venturini et al., 2015*[43]

Cosa rende un'analisi di *topic modeling* uno strumento utile per il monitoraggio di uno specifico ambito? L'osservazione delle relazioni *topic-keywords e topic-keyword* in chiave micro/macro

☞ Facilita letture critiche sui pattern di similarità tra topic;

☞ Migliora l'identificazione di concetti chiave dell'ambito specifico;

☞ Aiuta l'analista/osservatore ad identificare una o più narrazioni che permettano di conoscere buona parte del fenomeno;

☞ Facilita una lettura mirata dei documenti.

In *Termite : Visualization Techniques for Assessing Textual* Topic *Models (J.*Chuang et al., 2012) un progetto sull'identificazione degli argomenti trattati nei 372 IEEE InfoVis conference papers dal 1995 al 2010 la relazione tra topic e *keywords* trova espressione in una *bubble chart* [44] e in un *barchart*[45] orizzontale con tutte le *keywords* disposte in ordine di frequenza assoluta. Selezionare un singolo topic o una singola *keyword* consente di visualizzare nel dettaglio composizione e relazioni specifiche. (fig. ☞ 28)

43. Venturini, T., Bounegru, L., Jacomy, M. & Gray, J. How to Tell Stories with Networks: Exploring the Narrative Affordances of Graphs with the Iliad. datafied Soc. Stud. Cult. through data 1–13 (2015).

44. Tipo di grafico che mostra due o tre dimensioni di dati.

45. Tipo di grafico a barre la cui lunghezza è proporzionale al valore che rappresentano.

In altri esempi la condivisione di parole tra i topic, in parte evidente dalla matrice è enfatizzato da modelli visivi a rete (fig. ☞ 26) come nel caso di *Interactive Visual Exploration of* Topic *Models using Graphs* (Ronqvist et al., 2014) in cui dalla il modello evidenzia come

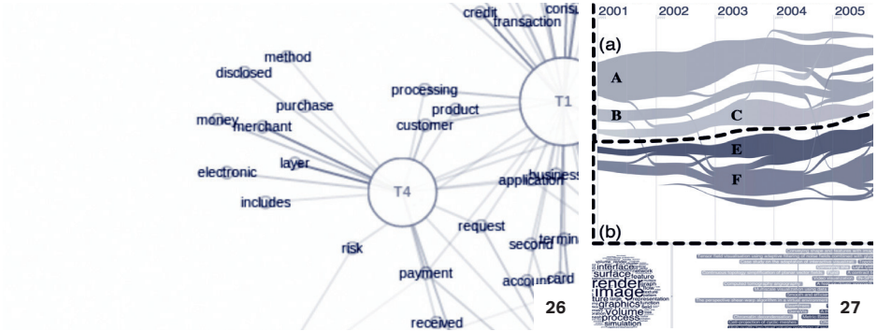
> *some terms are shared among* topic, *which hints at their ambiguity*
>
> — *S. Rönqvist et al., 2014*[46]

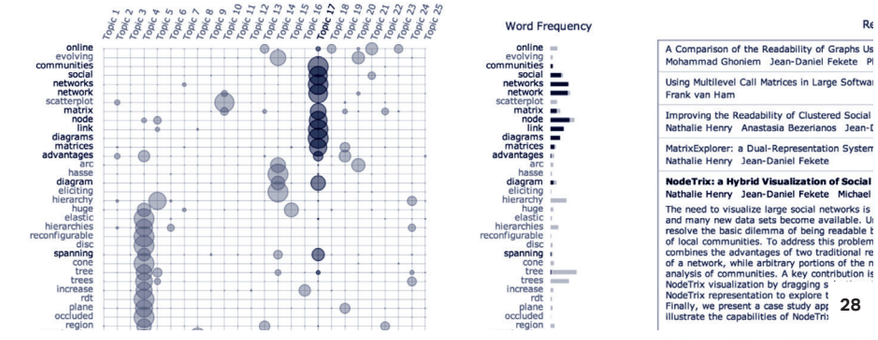
Questo genere di considerazione non è immediato se il modello visivo scelto è una *word cloud*[47] come nel caso di *Textflow* che aggrega le *keywords* tipiche di tutti i topic senza consentire una lettura in chiave macro/micro.
Pur non trattandosi di una visualizzazione di risultati di *topic modeling* il progetto *Trump Connection* (K. Albrecht, 2016) permette di esplorare 1500 entità connesse con Donald Trump sulla base di un set di dati raccolto da *Buzzfeed*[48] (fig. ☞ 29). L'aspetto interessante della visualizzazione interattiva è l'esasperazione del concetto di micro/macro lettura attraverso il passaggio da network a dettagliati dendrogrammi.

46. 1. Rönnqvist, S., Wang, X. & Sarlin, P. Interactive Visual Exploration of Topic Models using Graphs. (2014).

47. Modello visivo che raggruppa parole in una nuvola. La dimensione delle parole è proporzionale alla loro rilevanza.

48. Sito web d'informazione





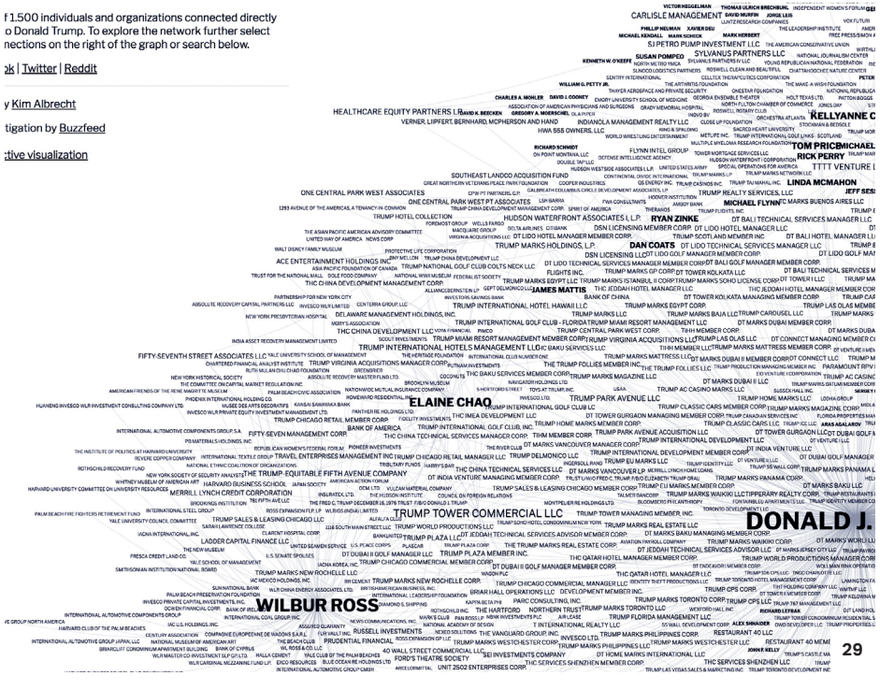


fig.26 S. Ronqvist et al., *Interactive Visual Exploration of* Topic *Models using Graph*s, 2014C

fig.27 W. Cui et al., *Textflow*, 2011. Elementi principali dell'interfaccia

fig.28 J. Chuang et al., *Termite*, 2012. Dettaglio dell'interfaccia con relazioni keyowrd/topic/documenti.

fig.29 K. Albrecht, *Trump Connections*, 2016. Network principale delle connessioni di D. Trump a persone. (http://trump.kimalbrecht.com/#11)

fig.30 Tassonomia per il *topic modeling* sulla base della bibliografia analizzata.

4.2 Una nuova combinazione

Poiché dalla revisione della letteratura non sembra esserci alcun testo che metta in luce lo stato dell'arte della visualizzazione dei risultati degli algoritmi di *topic modeling* e che mostri quali siano stati i modelli visivi più utilizzati negli anni per rappresentare questo genere di dato, tale organizzazione delle fonti è stata non solo uno strumento in grado di definire chiaramente lo stato dell'arte, ma anche un sistema utile per valutare l'efficienza del progetto *TopTom* (cap. 05) e indagare, con metodo, strade per possibili implementazioni future. (fig. 30)
Dall'analisi della bibliografia è emersa la mancanza di un esempio completo che mostri attraverso un'interfaccia dinamica dati relativi ai parametri enunciati in precedenza: contenuto (nello specifico contenuti anomali), gerarchia, relazioni e cronologia; *TopTom*, progetto innovativo di ricerca, ha permesso di sperimentare in questa direzione.

Non sapete che cosa sia, lo spazio! Credete che consista di due sole dimensioni. Io, invece, sono venuto ad annunciarvene una terza.

— Edwin Abbott, Flatlandia, 1882

5. Caso studio: Topic Tomographies

5.1 Il progetto

5.1.1 *Opidemic*

> *Opidemic, negli Usa la piaga dei farmaci che uccidono più del dolore*
>
> — *G. Tett, 4 marzo 2017*[49]

Ogni giorno circa 90 americani muoiono per overdose da oppioidi, assunti sotto forma di medicinali con regolare prescrizione o di eroina.
Questa tendenza è stata da tempo identificata dai media nazionali ed internazionali come *crisi degli oppiacei* o *oppidemia*.

49. http://www.ilsole24ore.com/art/commenti-e-idee/2017-03-04/opidemic-usa-piaga-farmaci-che-uccidono-piu-dolore-155653.shtml

Restringere severamente le possibilità di prescrizione di oppioidi da parte dei medici potrebbe porre fine a questo problema, ma le potenti lobby farmaceutiche sono ostili, per cui tra la totale liberalizzazione della vendita di oppiacei e l'opposta restrizione, un dettagliato monitoraggio della discussione online attraverso, per esempio, l'applicazione del *topic modeling*, consentirebbe a soggetti come governo, sanità e giustizia di identificare pattern di comportamento degli utenti ed agire di conseguenza nella promulgazione di leggi e nella gestione del fenomeno[50].

5.1.2 Una collaborazione tra discipline diverse

Il progetto di ricerca *TopTom* nasce dalla collaborazione tra *DensityDesign*[51] e il dipartimento torinese di ISI Foundation (Istituto per l'Interscambio Scientifico)[52] nel maggio 2017 con lo scopo di progettare un tool interattivo che visualizzi i risultati di un'analisi di *topic modeling*.

Team di progetto
Duilio Balsamo – data scientist
Paolo Bajardi – data scientist
Paolo Ciuccarelli – designer
Beatrice Gobbo – designer
Michele Mauri – designer
Andrè Panisson – data scientist
Riccardo Scalco – sviluppatore

Il team di ISI Foundation si è occupato dell'acquisizione dei dati e della *topic detection* attraverso l'uso di algoritmi supervisionati e non supervisionati.
Il team di *DensityDesign* si è invece occupato della progettazione del tool di visualizzazione.
Beatrice Gobbo si è dedicata alla ricerca bibliografica,

50. I dati relativi al tema degli Oppiacei sono solo un set di esempio, sia il sistema di algoritmi che la visualizzazione dati deve garantire la compatibilità con dati estratti da temi differenti con analogia di comportamento sulla rete.

51. Laboratorio di ricerca del dipartimento di Design del Politecnico di Milano. Dal 2004 si occupa di ricerca nell'ambito dell'information design e della data visualization per fenomeni sociali complessi.

52. Centro statale di ricerca scientifica con sede a Torino e New York. Dal 1983 si occupa prevalentemente di network e sistemi complessi, fisica matematica, statistica e quantistica.

all' elaborazione del concept, alla progettazione dell'architettura dell'informazione e dell'interfaccia. Riccardo Scalco si è occupato della programmazione del tool.

5.1.3 Uno strumento "semplice"

OBIETTIVO: analisi dati fenomeno specifico e individuazione variazione dei trends
FONTE DATI: reddit, twitter, gdelt
DEVICE: computer desktop
PIATTAFORMA: Google Chrome
INTERFACCIA: intuitiva e semplice
MODELLO VISIVO: noto

Modello visivo noto

Obiettivo del progetto è l'implementazione di una piattaforma in grado di analizzare dati che caratterizzano un generico fenomeno osservato, provenienti da diverse sorgenti (siti web, rss feed, blog, forum, social media e documenti in formato digitale), per individuare le variazioni nel tempo del fenomeno che indicano variazioni nello scenario in studio ed emergenza di trend.
Obiettivo dell'interfaccia è quella di introdurre l'utente al *corpus* preso in analisi, fornire strumenti per la sua esplorazione temporale, ed infine dare la possibilità di effettuare approfondimenti in un'ottica di lettura superficiale ed attenta.

Condizioni sine qua non

Durante il lancio del progetto sono stati elencati gli elementi fondamentali che devono essere presenti nel tool.

Il tool deve:

☞ Consentire di visualizzare le relazioni con le entità riconosciute nei documenti e le parole chiave identificate in modo non supervisionato, fornendo consapevolezza sullo scenario globale;

☞ Visualizzare ed esplorare dati aggregati su diverse scale temporali; consentire di fare un approfondimento su topic che presentano trend anomali, identificati precedentemente in maniera automatica.

Questa lista di elementi fondamentali combacia e completa la lista prodotta nel capitolo precedente come risultato dell'analisi della bibliografia. Essendo l'interfaccia pensata per essere fruita tramite un computer desktop, le interazioni principali saranno progettate per essere fruite via mouse e tastiera, comprendendo la possibilità di effettuare zoom su porzioni delle visualizzazioni per leggerne i dettagli, la possibilità di avere informazioni di dettaglio sui singoli elementi tramite pannelli e pop-up, la possibilità di definire filtri temporali per poter confrontare il comportamento del *corpus* in momenti differenti, la possibilità di effettuare approfondimenti visualizzando i singoli documenti (tweets, commenti su Reddit, singole notizie).

Diverse prospettive

ISI Foundation ha fornito un diagramma in grado di mostrare i processi principali alla base dell'estrazione e clusterizzazione dati in cui è mostrato il funzionamento di training dell'algoritmo attraverso le diverse fasi di *topic modeling*. Il processo di visualizzazione può essere espanso in diverse componenti.
La figura mostra come il processo progettuale lato

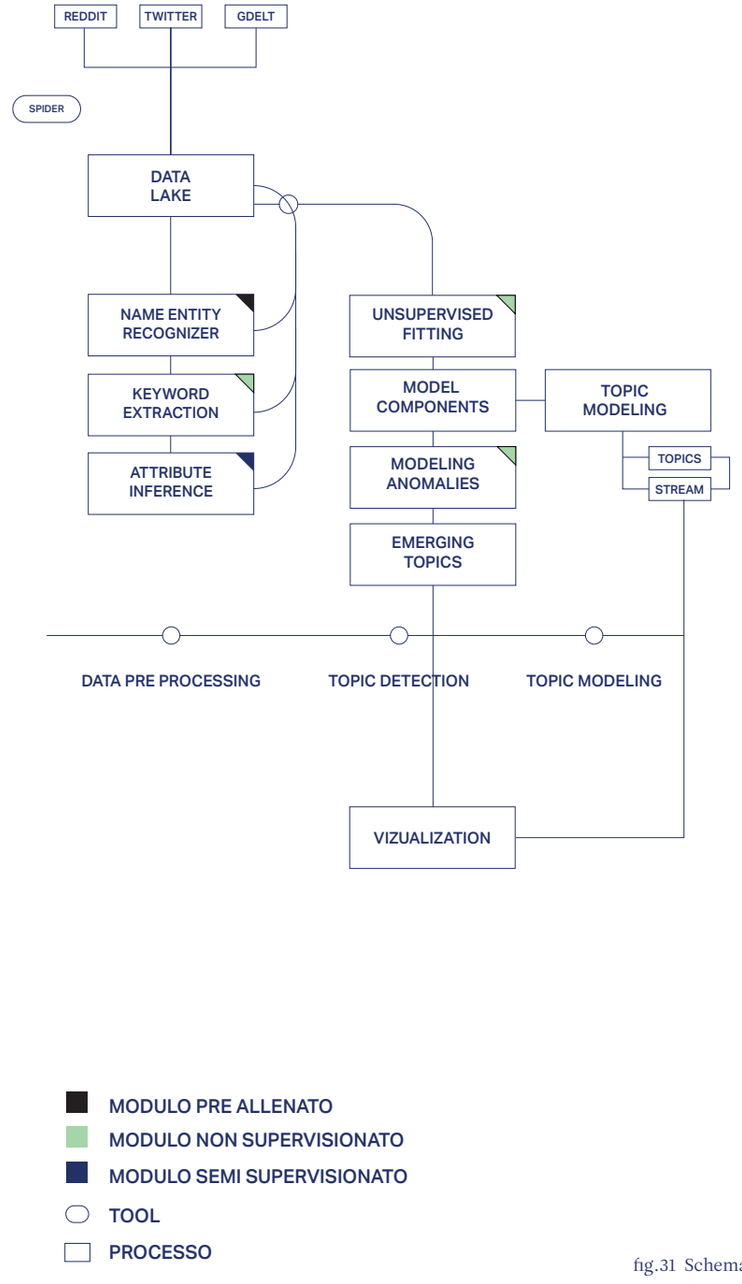


fig.31 Schema dell'architettura funzionale del sistema proposto da ISI Foundation in occasione del lancio del progetto.

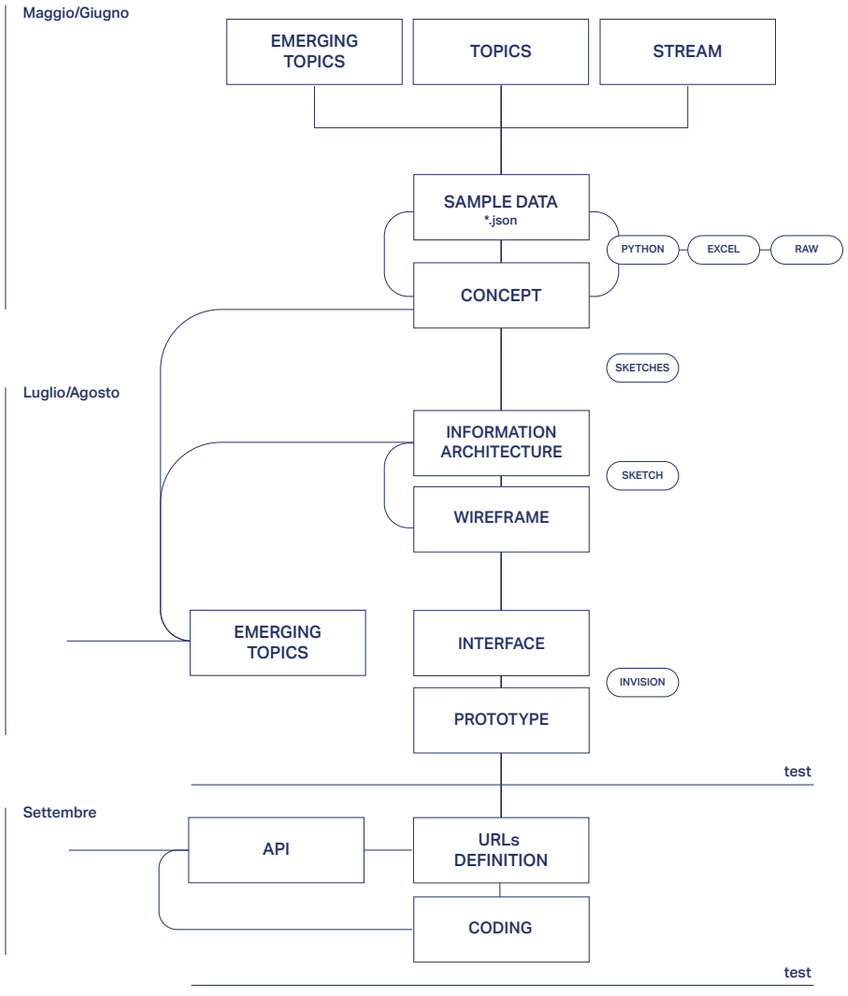


fig.32 Schema della task "visualization" esplosa in tutti i suoi passaggi, con riferimento temporale ai tempi di progetto.

design sia strettamente correlato al processo di estrazione e lavorazione dati protratto nel tempo.
Nel corso della progettazione, infatti, la sinergia di lavoro tra i diversi team ha permesso di apportare modifiche a dati e visualizzazione costantemente.
(fig. ☞ 31 e 32)

Una serie di algoritmi

Nella seguente sezione saranno elencati i principali processi che hanno consentito di definire il dataset sul quale il progetto è costruito.
Il team di ISI Foundation ha effettuato la *topic detection* partendo sulla base di quattro passaggi principali.
Ad una fase iniziale di *pre processing pipeline*, in cui sono state rimosse le *stop words*, identificati i nomi, le entità e i luoghi chiamata, ha seguito la fase di valutazione delle *keywords* identificate attraverso il calcolo del *tf-idf*. In seguito, ad ogni topic, identificato da 20 *keywords* con valori diversi, è stato assegnato un valore temporale e sono state identificate le anomalie presenti (improvvisi picchi negativi o positivi di valore).
(fig. ☞ 33)

PRE-PROCESSING

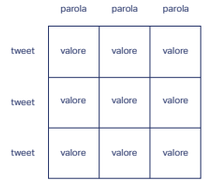
da documenti su uso e abuso di oppiacei in USA vengono estatte keywords che vengono poi filtrate

raccolta dati grezzi sulla base delle queries di ricerca ed eliminazione di stopwords

TF - IDF

ad ogni keyword del dizionario viene assegnato un valore calcolando la frequenza di ogni termine nei singoli documenti e nel corpus.

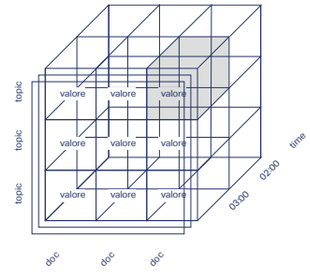
il valore delle keywords è pesato in relazione al documento di appartenenza. (vedi cap. 2)



NMF

ad ogni topic, caratterizzato da 20 keywords con diversi valori, viene assegnato un valore nel tempo e il grado di anomalia.

attraverso una struttura a matrice tridimensionale (tensore) si identificano topics e keywords per diversi intervalli temporali



UPGMA

i topic estratti vengono clulsterizzati a due a due a seconda di una matrice di similarità.

topic simili vengono accorpati a due a due fino ad ottenere un singolo topic molto generico

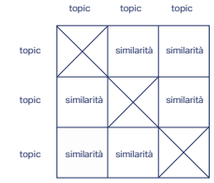


fig.33 Schema della serie di algoritmi usati dal team di ISI Foundation per effettuare la *topic detection*.

5.1.4 Il SuperUtente

> *TopTom è un "sensore" near-real-time rispetto a un fenomeno: ne fa vedere l'andamento, mette in fila le idee, mostra cosa è successo per esempio nella notte rispetto al* topic *che sto osservando"*
>
> — *SuperUtente*[53]

La ricerca sull'utente è stata svolta in un primo momento incontrando i principali utilizzatori del tool. Trattandosi di un gruppo ristretto i consigli e i feedback sono sempre molto specifici e indirizzati.
Il costante confronto con l'utente ha consentito di creare una rete di feedback - modifiche - test quasi simultanea. Gli esperti di dominio sono analisti che conoscono nel dettaglio l'argomento trattato nei dati raccolti ma hanno anche competenze da analisti.
Nel caso del set di riferimento con cui è stato progettato *TopTom* gli utenti sono molto esperti in tossicologia, legislazione sanitaria, medicina, farmacia e analisi dati.
I ricercatori di ISI Foundation, computer scientists esperti addetti alla costruzione e gestione dell'algoritmo, incarnano la figura di connessione necessaria tra gli esperti di dominio e i designer di *DensityDesign*.
Tipicamente all'esperto di dominio sono assegnate due tipologie di analisi:

tattica: puntuale su specifici termini chiave
strategica: profilazione di un fenomeno

TopTom lavora sul secondo task/obiettivo dell'utente che può essere descritto come

> *strategico e funzionale all'esplorazione del fenomeno e che potrebbe diventare, in futuro, uno strumento di previsione sulla base di pattern di comportamento.*
>
> — *SuperUtente*

53. L'esperto di dominio - super utente è un analista che conosce la tematica affrontata ed ha competenze nel campo della *data science*.

L'utente deve poter analizzare il topic in maniera funzionale all'esplorazione del fenomeno nella sua totalità, attraverso l'approfondimento di singoli elementi e la possibilità di selezionare diversi intervalli e granularità temporali.
Data la provenienza multi-settore degli utenti e considerando che non hanno nessun tipo di esperienza con avanzati sistemi di analisi dati, era necessario che *TopTom* fosse, innanzitutto semplice con funzioni simili agli strumenti già esistenti.

5.1.5 Design a tutto tondo

Un aspetto interessante del progetto è l'opportunità di vedere il designer come mediatore in occasione di diverse fasi di progetto.
La gestione della collaborazione simultanea di diverse figure professionali, con competenze e metodi progettuali diversi è essa stessa parte del processo di design.

> *Il designer della comunicazione, è dunque comparabile alla figura di un traduttore, in quanto, attraverso procedure di configurazione e trasferimento, svolge una attività di mediazione continua tra gli elementi di contesto e la diversità degli attori*
>
> *Il designer della comunicazione ha abilità specifiche e competenze trasversali che si attuano nell'interpretazione e nell'organizzazione dei contenuti, nel loro trasferimento da un contesto a un altro, nell'invenzione di nuovi interpretanti e abiti sociali, che rinnovano la nostra relazione con le cose.*
>
> — *G. Baule et al., 2016*[54]

54. Baule, G., Caratti, E., Design è traduzione. Il paradigma traduttivo per la cultura del progetto. «Design e traduzione»: un manifesto, (introduzione),Franco Angeli Editore, Milano, 2016

Nel caso specifico del progetto *TopTom*, è stato necessario comprendere a fondo la forma del risultato della analisi dei data scientists per poter individuare una metafora visiva calzante e in molte occasioni è stato chiaro come l'efficienza di funzionamento di un modello visivo fosse correlata non solo al contenuto, ma alla struttura stessa del dataset, così come l'efficienza di funzionamento del tool fosse garantita dall'efficacia simultanea di modello visivo e interfaccia.

TopTom sarebbe dovuta essere una visualizzazione di dati interattiva, inserita all'interno di un'interfaccia, che funzionasse attraverso uno stream di dati e potesse essere condivisa facilmente ad ogni livello di esplorazione attraverso le URLs. Queste quattro caratteristiche del progetto possono essere riassunto in quattro differenti aspetti del processo di design.
(fig. ☞ 34)

Design del dataset
Il modello visivo è costruito sulla base di dati strutturati ad hoc e l'interfaccia funziona perché il dato è costruito seguendo un metodo specifico.
Quando si progettano tool interattivi e non visualizzazioni statiche bisogna tenere in considerazione che il modello visivo dovrà adattarsi ai dati in tempo reale; per cui la struttura dei dati dovrà innanzitutto essere sempre la stessa e il modello visivo dovrà essere pensato in modo che anche una variazione repentina di un valore del dataset non ne infici la leggibilità.

Design del modello visivo
Ogni tipologia di dataset può essere rappresentata solo con determinati modelli visivi. Il modello visivo nasce dall'esigenza di voler visualizzare una o più

caratteristiche del dataset.

Design dell'interfaccia
Nei più performanti tool interattivi di visualizzazione dati interfaccia e modello visivo si fondono, fino a creare un unico sistema interattivo. La visualizzazione stessa fa parte del contesto con cui l'utente può interagire, avvicinandolo al dato, dando la possibilità di toccare il dato.

Design delle URLs
Poiché *TopTom* è pensato per essere un tool di monitoraggio di perturbazioni, era necessario che i principali utilizzatori potessero scambiarsi tramite condivisione di link viste aggiornate a specifici livelli di analisi.

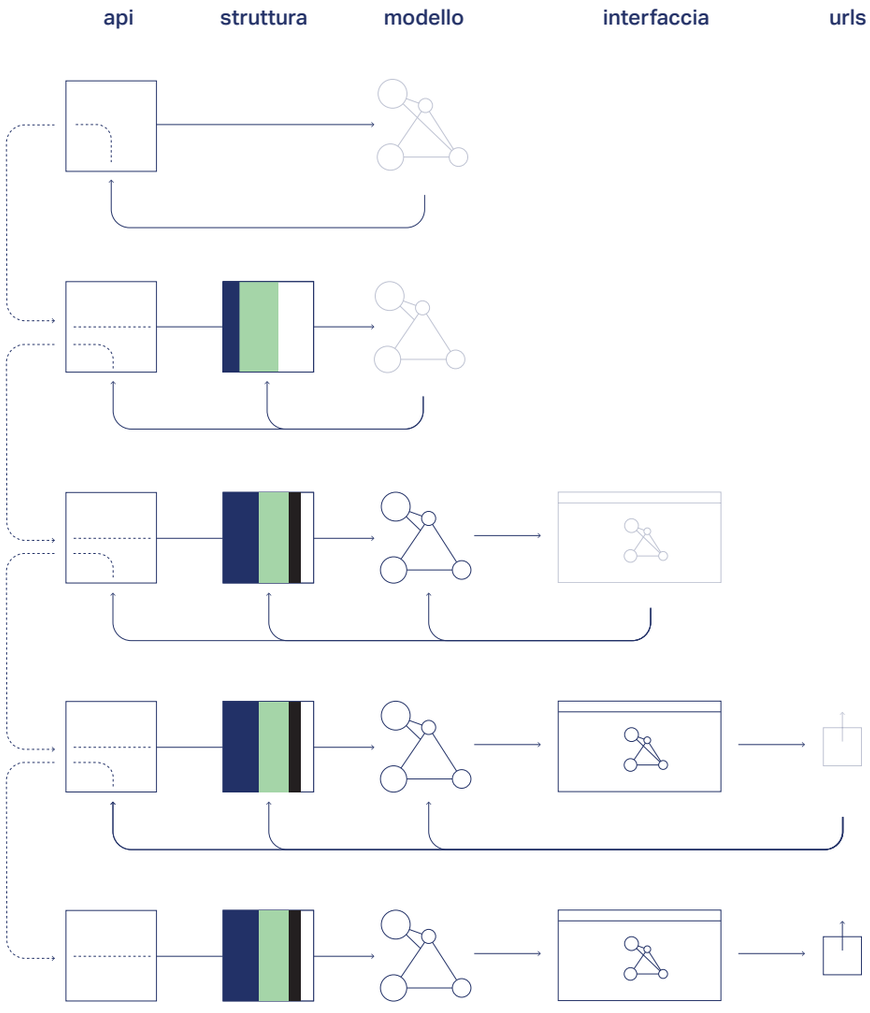


fig.34 Prima di arrivare alla definizione del modello visivo è stato necessario strutturare API specifiche. Una volta trovato il corretto modello visivo, non era scontato che funzionasse all'interno dell'interfaccia per cui sono stati rimaneggiati i dati. Analogamente è accaduto anche con la definizione delle URLs.

5.2 Dati gerarchicamente nidificati

Il formato file JSON (*JavaScript Object Notation*) è un formato adatto all'interscambio di dati fra applicazioni client-server, basato sul linguaggio Javascript.
JSON è un formato di testo completamente indipendente dal linguaggio di programmazione, ma utilizza convenzioni conosciute dai programmatori per molti altri linguaggi. Questa caratteristica fa di JSON un formato ideale per lo scambio di dati.
Nel mondo dell'informatica e della programmazione è considerato un *formato lightweight* perché riesce a sintetizzare grandi quantità di dati. Tuttavia, non si può affermare lo stesso per il mondo del design, in quanto è un formato dati complesso che richiede competenze specifiche per essere convertito in una forma leggibile e rappresentabile. Un file JSON è basato su due strutture principali: un insieme di coppie nome-valore ed un elenco ordinato di valori che può prendere queste forme:

Oggetto: serie ordinata di nomi/valori

```
{“string” : value,“string” : value}
```

Array: una raccolta ordinata di valori

```
[“value”,value]
```

Valore: può essere una stringa tra virgolette, un numero, vero, falso, nullo, un oggetto o un *array*.

Risale a Maggio 2017 il primo file JSON contenente un esempio di dati che ha permesso di iniziare ad esplorare il tema del *topic modeling* attraverso dati reali. (fig. ☞ 35, 36 e 37)
La prima struttura proposta aveva questa forma: un *oggetto* JSON con tre proprietà : nodi, intervalli temporali e dendrogramma.

```
{
“nodes”: { … },
“intervals” : [ … ],
“dendrogram”: { … }
}
```

La prima difficoltà incontrata è stata quella di comprendere a fondo la struttura del dataset di esempio proposto.
Ogni *nodo*, con il suo ID, definisce un topic.
Il campo *intervals* specifica il periodo di analisi di quel dataset o chiamata.
Il campo *dendrogram* definisce il livello di aggregazione gerarchica a cui si vuole osservare la clusterizzazione di topic.
Il dendrogramma (fig. ☞ 38) è un diffusissimo modello visivo per la rappresentazione del *coefficiente di similarità*[55], usato soprattutto in statistica, informatica e *data science*. Nell’asse delle ascisse esprime la distanza logica dei cluster secondo una specifica metrica definita, mentre sull’asse delle ordinate indica il livello gerarchico di aggregazione, solitamente compreso tra 0 e 1.
Nella figura 38 è mostrato un primo esempio di dendrogramma che si riferisce alla clusterizzazione gerarchica di topic durante una giornata composta da 24 ore.
Il livello di clusterizzazione con il coefficiente più alto (1) è il livello più alto del dendrogramma ovvero il cluster di topic più generico, mentre il livello con il coefficiente più basso (0) sarà il livello più basso del dendrogramma che definisce tutti i singoli nodi, chiamati anche *foglie* o *figli*.
Il livello 1 di aggregazione permette di visualizzare il topic più generale di un determinato frame temporale, mentre, andando gradualmente verso il livello di aggregazione 0, quindi verso il basso, si possono osservare cluster sempre più dettagliati fino a raggiungere il livello 0 in cui sono presenti ed identificabili tutti i topic singoli la cui quantità varia in maniera arbitraria, in relazione all’intervallo temporale scelto. Ad esempio, un intervallo temporale di 24 ore diviso in segmenti di un’ora, a livello 0 del dendrogramma, presenterà 5 (numero arbitrario, variabile a seconda del frame temporale) topic specifici per ogni ora per un totale di 120 topic singoli.

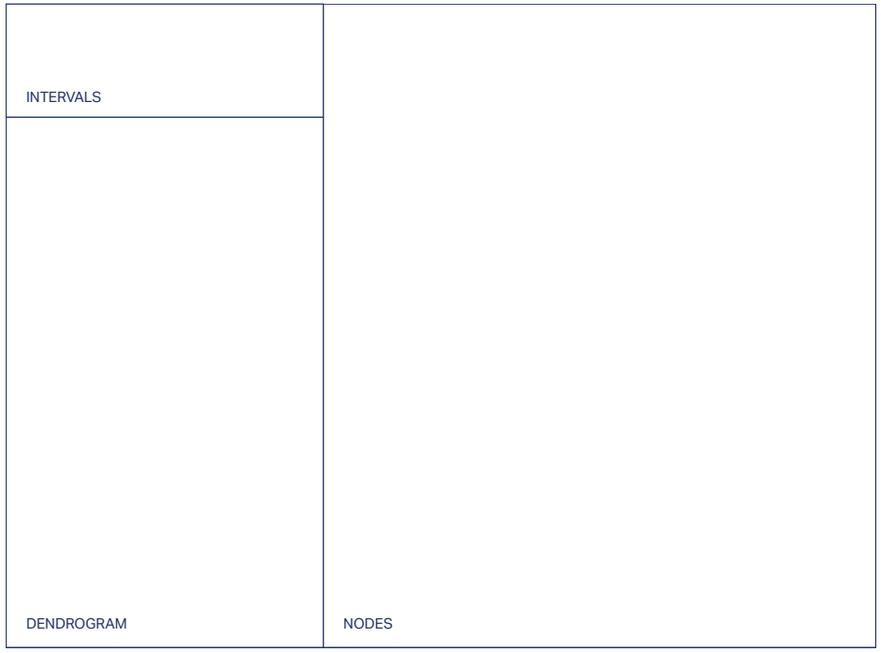
55. Coefficiente che indica la similarità tra due punti (nel caso del *topic modeling*, tra due nodi). Il coefficiente di similarità è calcolato come se fosse una distanza,

Per questo motivo chiameremo *topic* tutti i topic singoli di livello 0 e *cluster* tutti i topic emersi dalla clusterizzazione che raggruppa ad ogni livello di dettaglio i due topic (o cluster) più simili.

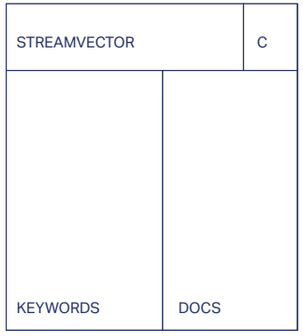
Nel dettaglio, l’*array* con l’id del *nodo*, contiene quattro oggetti, parametri fondamentali per definire il nodo, ovvero il topic.

```
{
“nodes”: {
“id_node”: {
"common_words": [ … ],
“stream_vector”: [ … ],
“topic_documents”: [...]
“children” : [ ];
},
```

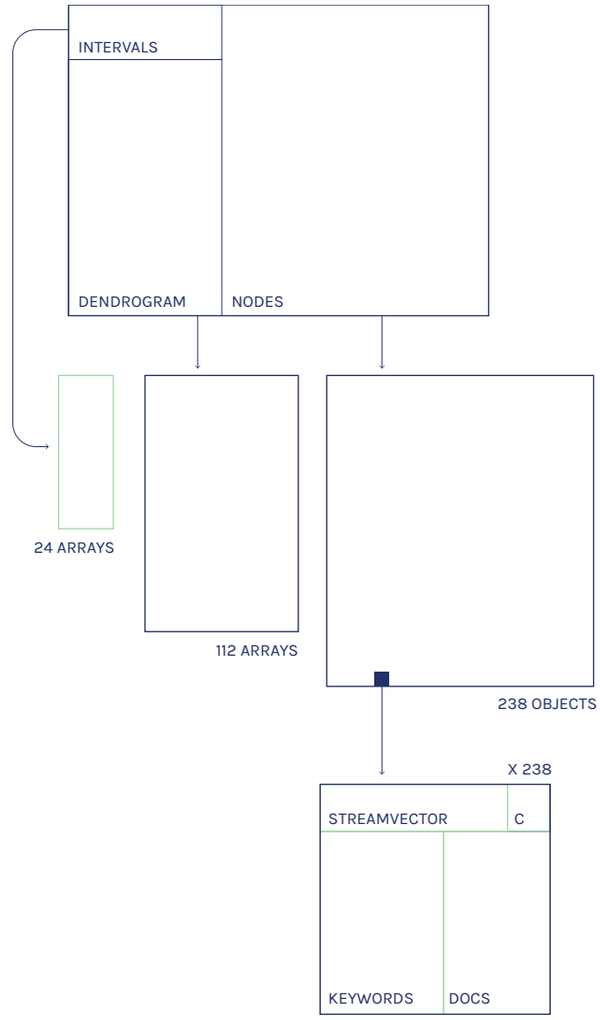
Il campo *common words* è un elenco ordinato di *keywords*



35



36



37

fig.35 `{“nodes”: {…},“intervals” : [ … ],“dendrogram”: { … }},.`
fig.36 Dettaglio dell’array del nodo di maggio 2017

fig.37 Rappresentazione esplosa del dataset di maggio 2017.

LEGENDA

Topic

Cluster

Livello
di aggregazione

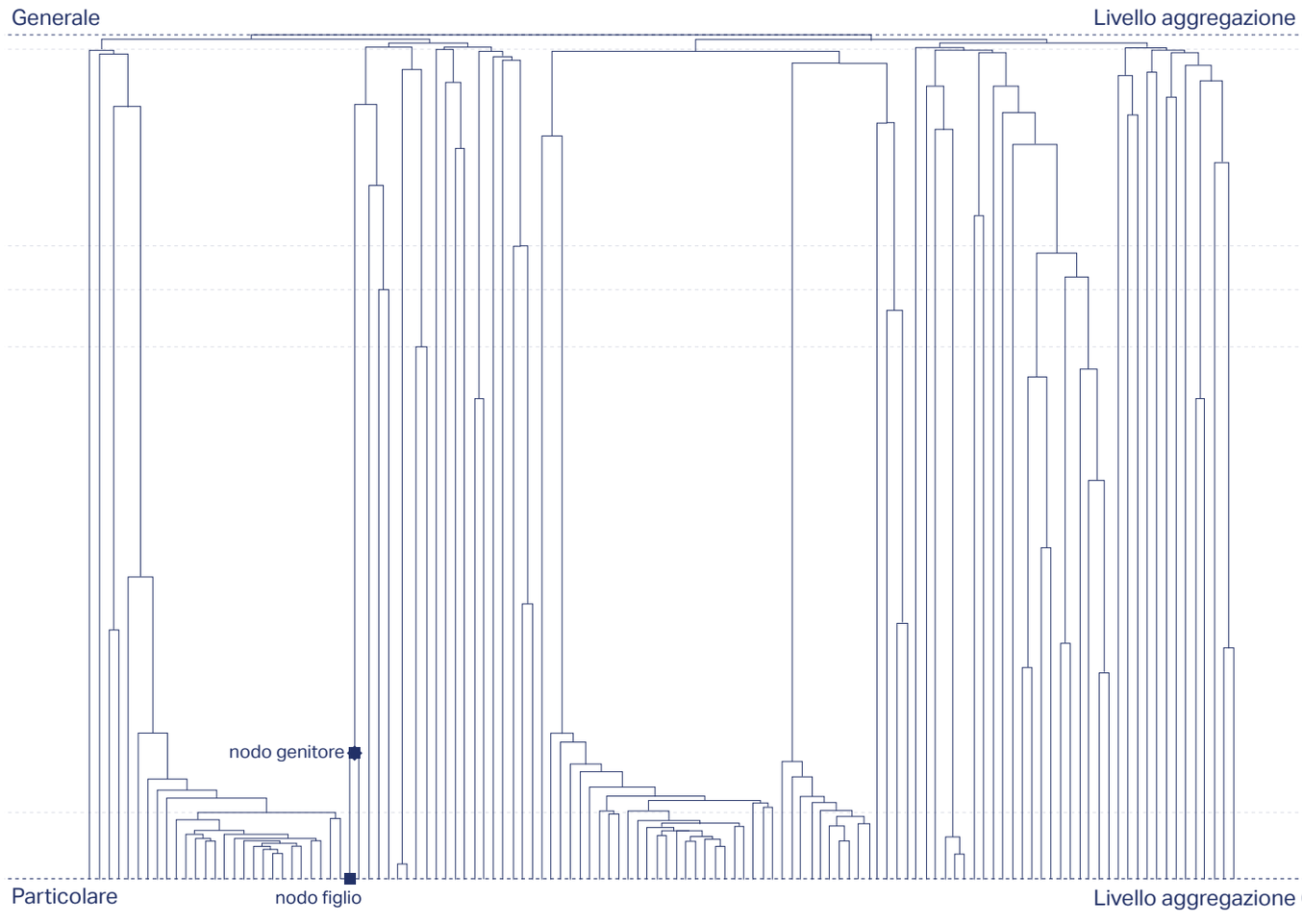


fig.38 Il dendrogramma, modello visivo usato per la rappresentazione di strutture gerarchiche.

Ad ogni *keywords* è associato un valore che indica la rilevanza di quella *keywords* all'interno di quel topic.
L'*array stream vector* definisce il valore di un topic negli intervalli temporali definiti nell'*array intervals*.
L'*array topic documents* definisce un numero arbitrario di documenti caratteristici di quel topic, anch'essi, come le *keywords* ordinati per rilevanza.
Ed infine l'*array children* definisce gli id dei figli diretti di quel nodo, ovvero i topic più vicini (simili) in fatto di *keywords*. Questo parametro che caratterizza ogni singolo nodo è alla base del dendrogramma mostrato in precedenza.

La struttura gerarchica

La clusterizzazione permette di definire i *children* di ogni nodo tenendo in considerazione la distanza o similarità tra i singoli nodi.
Il metodo *UPGMA* (Sokal R. e C. Michener, 1958) è comunemente rappresentato da un dendrogramma che riflette la struttura presente in una matrice a coppie di valori. Ad ogni step i topic o cluster più vicini sono combinati in un livello superiore di clusterizzazione.
Prendiamo ad esempio una *matrice con diagonale nulla*[56] 3x3. (fig. 👉 39)
a, *b* e *c* sono topic, i valori interni rappresentano la differenza tra di due topic, ovvero la loro distanza δ.
Identificata la prima sotto-matrice come coppia di valori più simili bisogna adesso posizionare i topic all'interno di un dendrogramma.
I topic *b* e *c*, poiché più simili si uniscono subito creando un cluster ad altezza *u*.
L'altezza *u* è calcolabile matematicamente in questo modo

```
hp : δ (b,u) = δ (c,u) = D1(b,c)/2 = 12/2 = 6
```

(fig. 👉 40).
In questo caso si può dire che i topic b e c sono figli del

56. Una matrice con diagonale nulla è il risultato dell'incrocio della stessa lista di valori. Quando i due valori omologhi si incrociano il risultato è zero.

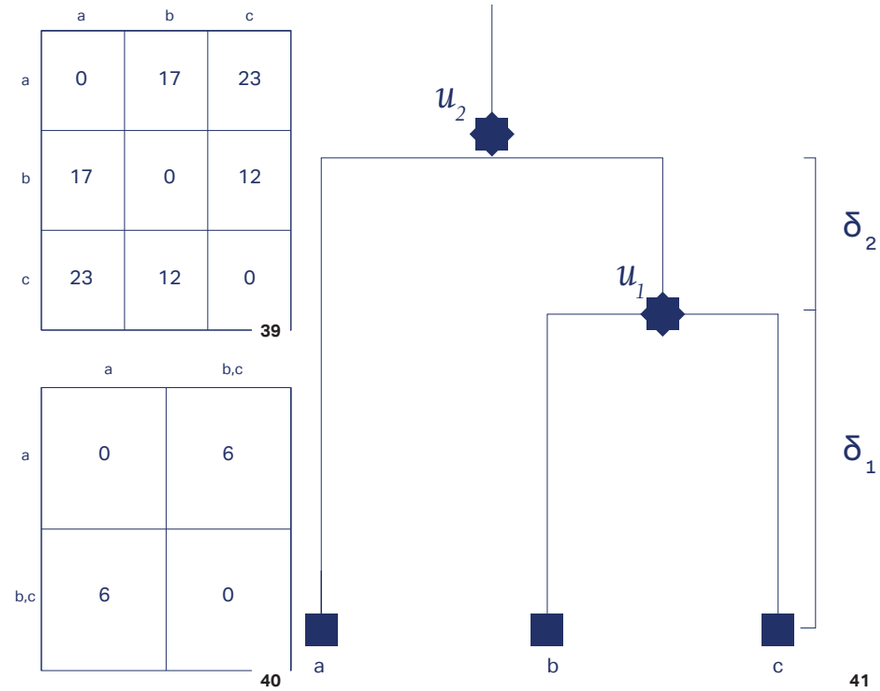


fig.39 Matrice sparsa mxn alla base della clusterizzazione UPGMA.
fig.40 Matrice risultante della clusterizzazione.
fig.41 Rappresentazione della clusterizzazione di due topic (b e con il metodo di clusterizzazione UPGMA..

cluster (b,c). (fig. ☞ 40 e 41)

Infatti a partire dalla versione del JSON di Giugno 2017 è stato aggiunto il campo *topic distances* che è una traduzione dei valori presenti in una matrice sparsa di dimensioni mxn.

Giugno 2017

```
{
“nodes”: { … },
“intervals” : [ … ],
“dendrogram”: { … }
“topic_distances”: [ … ]
}
```

Per comprendere meglio il concetto di similarità tra topic si può visualizzare attraverso un grafo che mette in evidenza la distanza tra i diversi topic. Più i nodi sono vicini più sono simili e quindi prossimi alla clusterizzazione.

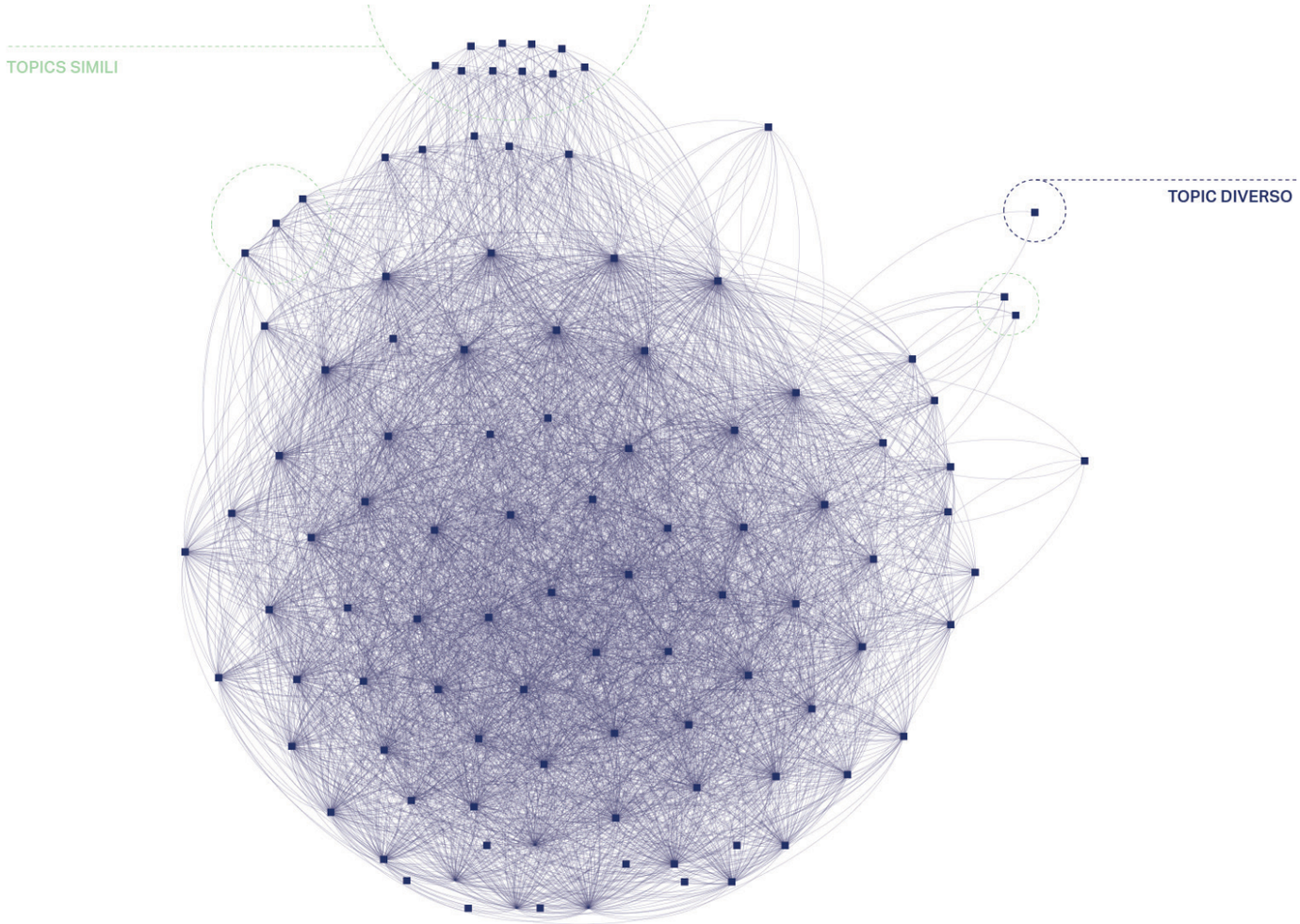


fig.42 I topic simili sono più vicini tra loro, pronti a formare un cluster.

La novità dell'anomalia

Una novità dal punto di vista algoritmico è la definizione del campo *anomalia*.
Già in passato alcuni studi prevedevano la possibilità di definire un campo del dataset in cui fossero presenti indicazioni relative agli aspetti più particolari di quel topic. Le anomalie sono definite come topic emergenti.

In *TopTom*, sono state definite due tipologie di anomalia

- *Volume decrease*: un'importante diminuzione della quantità di documenti che trattano l'argomento compreso in quel topic;

- *Volume increase*: un importante aumento della quantità di documenti che trattano l'argomento compreso in quel topic.

Il risultato della fase di *anomalies detection*, effettuata dal team di ISI Foundation, è stata una lista di anomalie in cui ogni elemento era definito da

- *Tipo* di anomalia (volume increase o volume decrease);

- *Livello* dell'anomalia;

- Topic in cui l'anomalia è stata identificata;

- *Span temporale* in cui l'anomalia è stata identificata.

Nella visualizzazione alla pagina successiva, è mostrato come nell'arco dei mesi di progetto il dataset abbia subito per lo più modifiche a Giugno2017, che coincide con la progettazione del modello visivo e

dell'interfaccia, e di carattere radicale quando, ad agosto 2017, è cominciata la parte di programmazione.
Entrambe le figure proposte (fig. 27 e 28) mostrano come il modello visivo a *treemap*[57] sia efficace per visualizzare gli *array* e gli oggetti che compongono un JSON.
Questo processo di visualizzazione pura dei dati si è rivelato molto utili ai fini della comprensione della struttura del dataset. (fig. 43)

57. Un metodo per mostrare dati gerarchici usando rettangoli innestati.

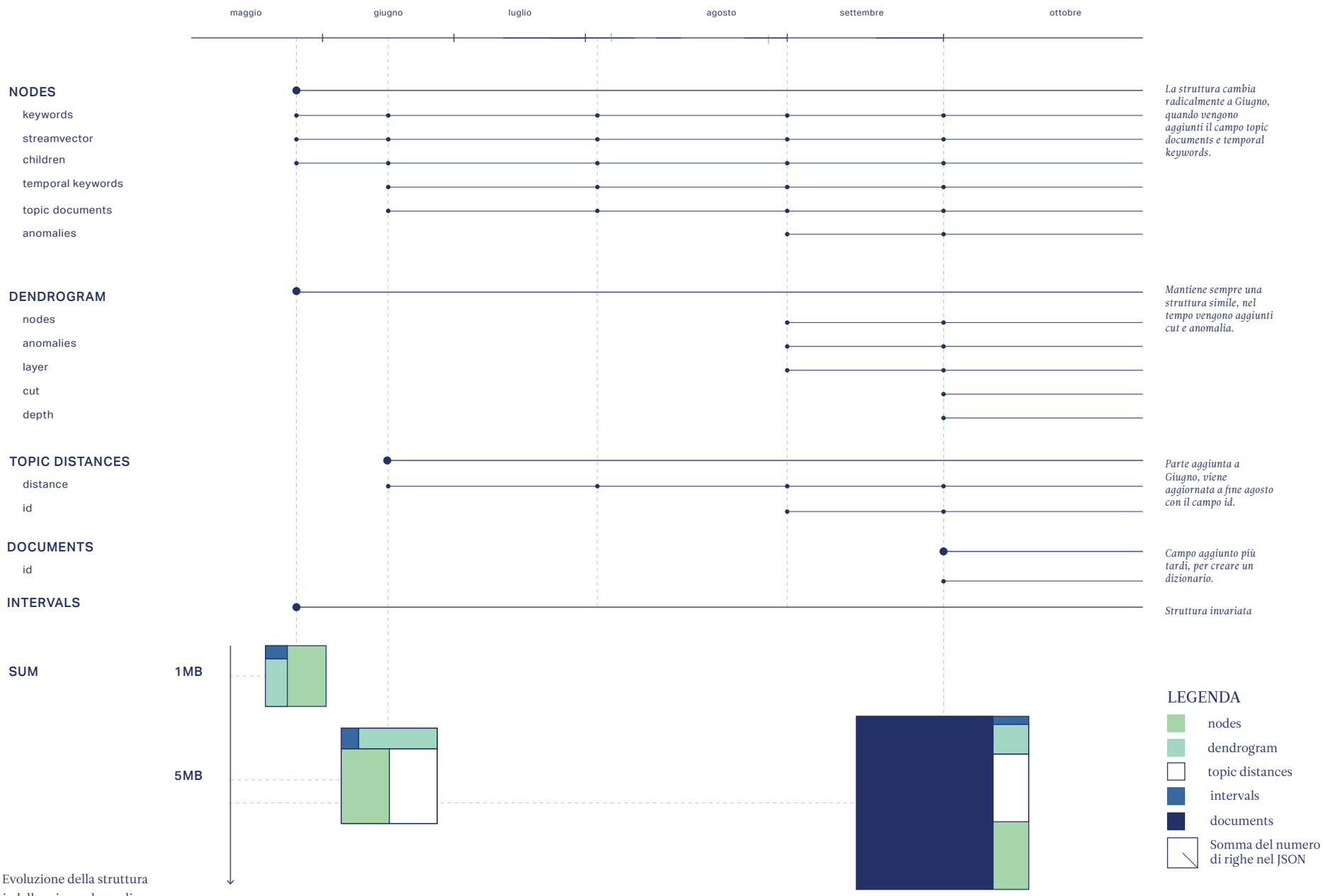


fig.43 Evoluzione della struttura dei dati, dalla prima release di giugno ad ottobre 2017.

5.3 Oltre la dashboard

Uno dei grandi limiti dei progetti elencati precedentemente (cap. ☞ 04) sta nella scarsa considerazione dell'interfaccia e del modello visivo come un'unica entità.
L'utente, infatti, deve potersi muovere tra diverse schermate per osservare il fenomeno da differenti angolazioni e, al contempo, vedere il tutto in maniera aggregata e dettagliata.

> *It is necessary to create different perspectives on the database, giving users the ability to shift from one to the other in a coherent way.*
>
> — *M. Mauri et al., 2013*[58]

Con *TopTom* la sfida iniziale è stata quella di creare un tool che non si limitasse ad essere una dashboard, come nel caso di *Hierarchical Topics* (W. Dou et al., 2013), *TextFlow* (W. Cui et al., 2011) o *#FluxFlow* (J. Zhao et al., 2014) ma un sistema di interazione complesso a più viste, sempre più dettagliate, in ottica micro/macro letture. Innanzitutto era necessario comprendere quali dovessero essere le viste principali e come organizzare le connessioni tra le viste tenendo in considerazione gli aspetti principali della *topic detection* di ISI Foundation: *topic modeling con temporal hierarchical clustering ed emerging topics.*

58. Mauri, M. et al. (2013) 'Weaving data, slicing views: A design approach to creating visual access for digital archival collections', ACM International Conference Proceeding Series CHItaly, pp. 1–8.

5.3.1 Task di interazione

TopTom nasce con l'intenzione di aiutare un gruppo di esperti di dominio a monitorare la diffusione di argomenti sui media, come se fosse un sensore quasi in tempo reale.
In linea generale, in un'ottica di macro-micro esplorazione (cap. ☞ 04), l'utente deve poter passare da una visione aggregata delle informazioni alla lettura dei documenti che caratterizzano un determinato topic, sia di un esteso livello temporale che di ad un intervallo specifico. In base alle necessità dell'utente è emerso come la funzione primaria del tool dovesse essere quello di:

☞ Visualizzare i topic anomali e perturbati;

☞ Osservare l'evoluzione dell'anomalia nel tempo;

☞ Relazionare la perturbazione ad altre sfaccettature del fenomeno;

Quattro filtri di informazione

L'utente esperto deve poter osservare l'anomalia a diversi livelli di dettaglio e granularità temporale, sia a confronto con altri topic anomali che in un contesto privato delle perturbazioni, per poterne identificare possibili cause ed effetti. Prendendo in considerazione i parametri definiti al capitolo 4, emerge come l'osservazione del fenomeno si coniughi sulla base delle loro possibili combinazioni. In questo modo il controllo da parte dell'utente di *gerarchia*, *cronologia*, *contenuto* e *relazioni* consente di monitorare ed analizzare il

fenomeno mostrandolo da angolazioni diverse ma senza trarre alcuna automatica conclusione, compito ancora affidato all'esperto di dominio; la struttura dell'architettura del sistema, osservabile sia in una rappresentazione generale (schema in alto) che in una più dettagliata (fig. ☞ 44) , non è altro che l'esplosione organizzata delle necessità dell'utente. In figura 30 emerge come la possibilità di vedere le anomalie sia un livello posto sopra ad ogni modello, che metta in risalto solo ed esclusivamente i contenuti anomali. *Trait d'union* tra la versione anomala e la versione neutra è la condivisione degli stessi parametri di analisi che consentono quindi un equo confronto tra le versioni. (fig. ☞ 45)

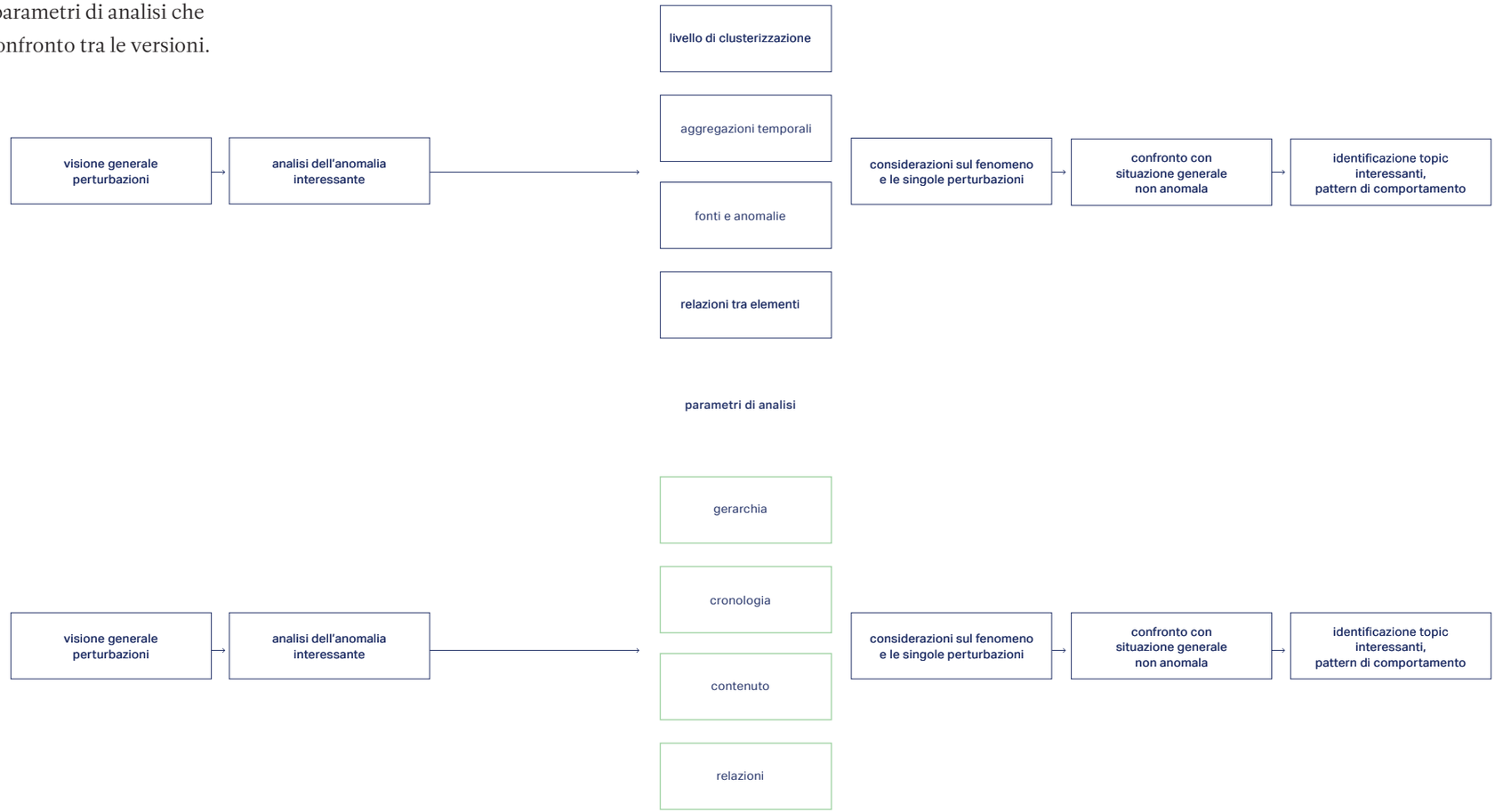


fig.44 Struttura primaria dell'architettura dell'informazione sulla base delle richieste dell'utente e della tassonomia del capitolo 4.

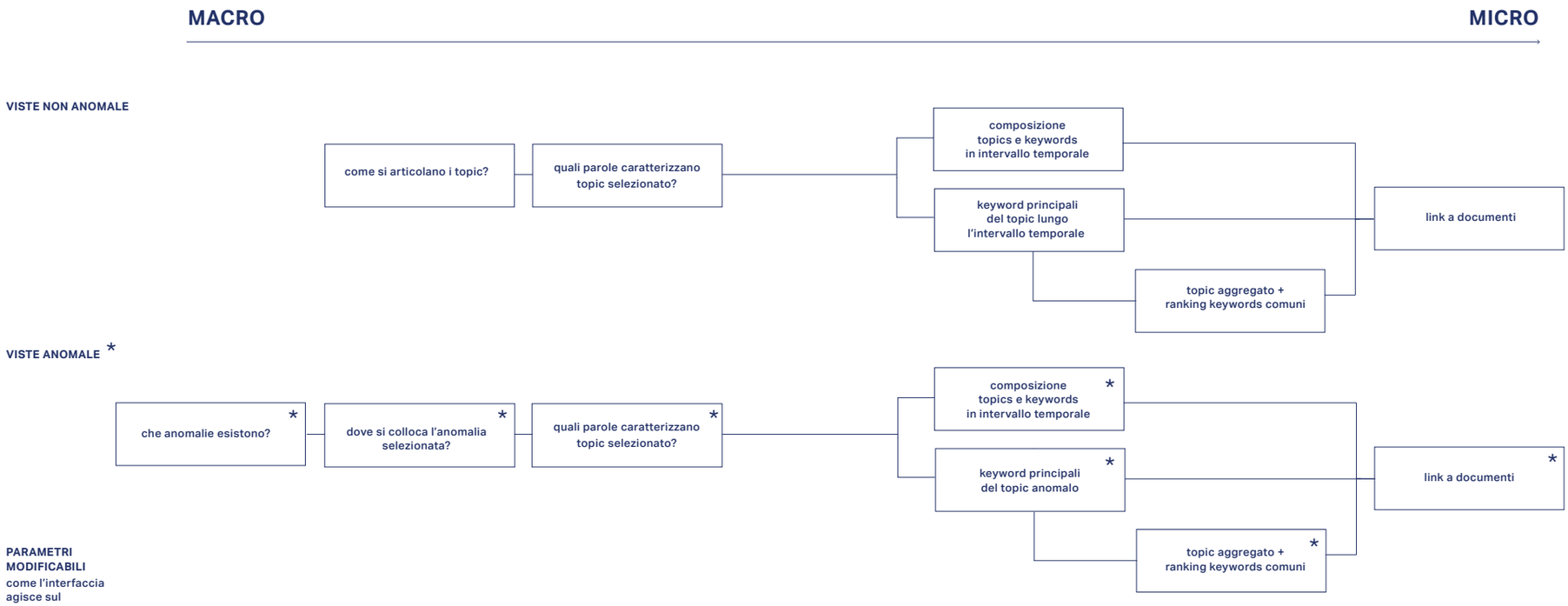


fig.45 Struttura esplosa dell'architettura dell'informazione. Gli stessi parametri possono essere applicati a una vista anomala o una vista "non anomala".

5.3.2 Metafore visive

Come rappresentare dati complessi forniti da ISI Foundation in una forma leggibile? Era necessario comprendere non solo la struttura del dato in sé, ma anche il suo comportamento. Prima di arrivare al modello visivo definitivo sono stati fatti diversi tentativi: uno dei tentativi che meglio ha permesso di comprendere la struttura del dataset e rappresentarlo è stata la *heatmap*[59]. Come già detto in precedenza, il primo risultato della *topic detection* di esempio fornito da ISI Foundation era riferito ad una giornata di 24 ore, per cui, per quella giornata, sono stati identificati 5 topic all'ora per un totale di 120 topic (dal topic numero 0 al topic numero 119). La prima *heatmap* rappresenta il caso di granularità più dettagliato della giornata; ovvero l'osservazione della struttura gerarchica alla sua base, con tutte le *foglie*.

Per identificare come i topic si associassero gli uni con gli altri salendo lungo il dendrogramma sono stati colorati i componenti della *heatmap* in base al cluster di appartenenza. A partire da questo modello si è poi fatto un tentativo rappresentando il dataset di esempio con *bumpchart*[60]. Il modello è stato considerato troppo confusionario e visivamente rumoroso ma ha permesso di identificare lo *streamgraph* come modello migliore, anche perché tra i casi studio citati uno dei modelli visivi più usati per visualizzare i risultati di *topic modeling* è lo *streamgraph*. Il primo modello di *streamgraph* risale al 1999, quando Susan Havre del Battelle Pacific Northwest DivIsion, Richland, USA pubblica uno studio sull'anallsi semantica dei discorsi di Fidel castro degli ultimi 40 anni. Affidandosi alla metafora del fiume

> *ThemeRiver simplifies the user's task of tracking individual themes through time by providing a continuous "flow" from one time point to the next. The horizontal flow of the river represents the flow of time. Each vertical section of the river corresponds to an ordered time slice.*
>
> — *S. Havre, 1999*

59. Modello visivo usato per mostrare variazioni di valore all'interno di una matrice a doppia entrata.

60. Modello visivo simile allo *streamgraph* utile per vedere variazioni nel tempo di diverse categorie. Rispetto allo *streamgraph* la *bumpchart* disponne secondo ranking ascendente o discendente sull'asse y ogni flusso rispetto ad ogni valore sull'asse x.

Nel 1999, su Infovis, viene presentato il progetto *ThemeRiver: In Search of Trends, Patterns, and Relationships* in cui viene sfruttata la metafora del fiume per rappresentare dei temi che si evolvono nel tempo.
Come già citato nel capitolo 4, la prima applicazione nel campo del design dello stesso modello visuale risale al 2008, quando sul New York Times appare *Ebb and Flow of Movies: Box Office Receipts Over Past 20 Years* di Lee Byron & Martin Wattenberg , due fig.ure professionali ibride, a cavallo tra *dataart* e *data science*.

> *A streamgraph (fig. ☞46), or stream graph, is a type of stacked area graph which is displaced around a central axis, resulting in a flowing, organic shape.*
>
> — *L. Byron e M. Wattenberg, 2008*

Lo *streamgraph*, così come lo *stacked barchart*[61], serve per visualizzare l'evoluzione e variazione di categorie nel tempo.
L'aspetto interessante della ricerca di Havre è la *metafora naturale / biologica alla base del modello*, gli argomenti si evolvono fluttuando orizzontalmente nel tempo come se fossero correnti di un fiume.
Tuttavia lo *streamgraph* non era l'unico modello visivo possibile per rappresentare il tempo. Sono stati fatti esperimenti con *scatterplot*[62], *bumpchart* ed *heatmap*, ma alla fine, è stato scelto lo *streamgraph* come migliore e più performante.
Per questo motivo anche nel caso di *TopTom* è stato scelto un approccio di questo tipo, partendo dal

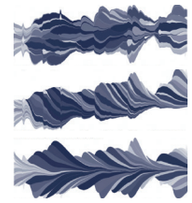


fig.46 *Streamgraph* di Byron e Wattenberg.

61. Modello visivo costituito da più insiemi di dati uno sopra l'altro al fine di mostrare come la categoria più grande è suddivisa in categorie più piccole e le loro relazioni con il totale.

62. Modello visivo che consente di visualizzare i dati in uno spazio x,y secondo la logica della dispersione.

concetto di fiume/flusso di Harve e cercando di ampliare la metafora in altre direzioni.

Sfogliare uno streamgraph

> *Escaping Flatland is the essential task of envIsioning information*
>
> — *E. Tufte, 1990*

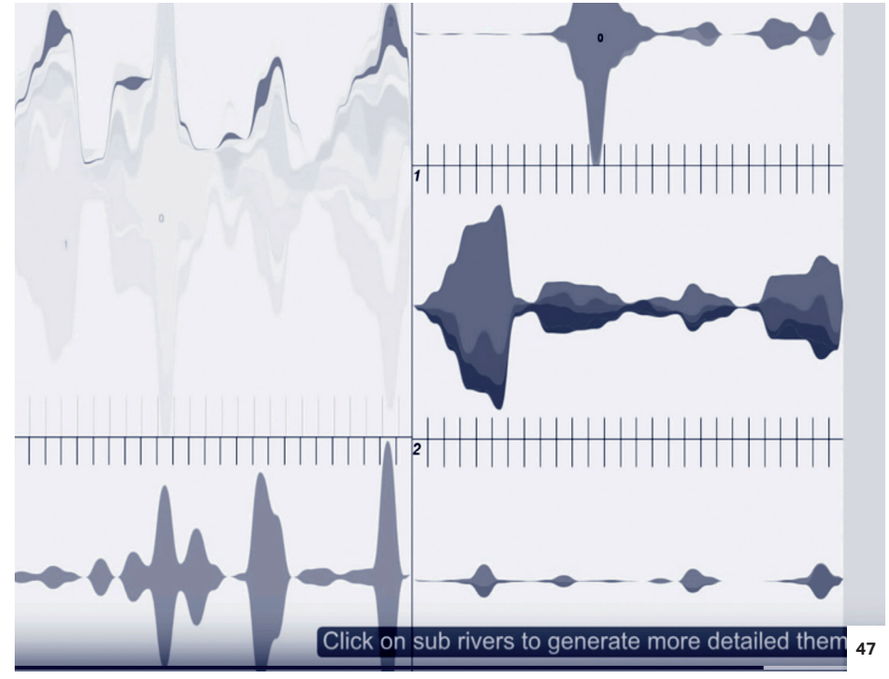
> *Elle seule pennet de répondre aux questions précédentes. Dans toute autre construction on ne voit que la feuille, à la rigueur la branche. Mais l'arbre est invIsible. X et Y sont les dimensions orthogonales du tableau. Z est la variation d'energie lumineuse en chaque point significatif du tableau. Cette' variation n'est obtenue que par la taille ou la valeur.*
>
> — *J. Bertin, 1967*[63]

Quando si ha a che fare con visualizzazioni interattive si tende ad evitare la rappresentazione tridimensionale perché difficilmente usabile da utenti abituati a interfacciarsi con strumenti come il *Notebook2* di IBM. Tuttavia, un approccio tridimensionale in fase di elaborazione del concept può essere di grande aiuto.

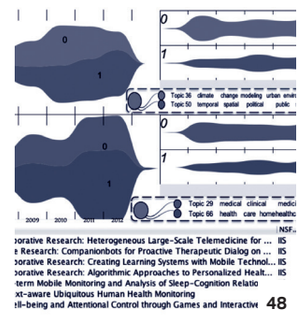
La caratteristica principale dell'algoritmo di *topic modeling* usato per *TopTom* è la presenza della profondità di dettaglio definita dal dendrogramma come nel caso di *Hierarchical Topics*.

> *Hierarchical Topics provides a scalable solution that allows iterative analysis of document collections with a large number of* topic *and further support the exploration of temporal evolution of those* topic *in a hierarchical fashion.*
>
> — *W.Dou et al., 2013*

63. Bertin, J. and Les, C. (2009) 'Sémiologie graphique', Image (Rochester, N.Y.), pp. 7–10. doi: 10.1037/023518.



47



48

49

fig.47 Sdoppiamento multiplo dell'interfaccia in *Hierarchical Topics* (W. Dou et al., 2013).
fig.48 Possibilità di annotazione in *Hierarchical Topics*.
fig.49 Rappresentazione del dendrogramma in *Hierarchical Topics*

Il livello più generico di topic è il livello 1, quello più dettagliato il livello 0 con n topic.
Se si provasse ad osservare uno *streamgraph* tridimensionale posizionato nello spazio a livello di profondità 1, la parte visibile sarebbe l'evoluzione temporale di un grande topic generico durante tutta la durata dell'intervallo temporale scelto, quello che, analogamente in due dimensioni, appare come *landing view* nell'interfaccia di *Hierarchical Topics* (W. Dou et al.. 2013).
La presenza di un solo topic che caratterizza tutto l'intervallo temporale è tuttavia molto generica e necessita di un'osservazione più dettagliata delle sue componenti per permettere un'analisi reale dell'argomento.

Osservare un dendrogramma a diversi tagli è paragonabile ad una sezione dello *streamgraph* gradualmente sempre più lontana dal punto di osservazione.
(fig. ☞ 51)
Un comportamento analogo è quello del macchinario usato per fare la *Tomografia Computerizzata* (TAC) (fig. ☞ 50) in ambito medico. La TAC è una metodica diagnostica per immagini, che consente di riprodurre sezioni o strati corporei del paziente ed effettuare elaborazioni tridimensionali. La parte superiore del cranio è il livello 1 del dendrogramma, la parte più vicina al collo è il livello 0. Questo genere di metafora consente all'utente di scegliere e sperimentare a che livello di dettaglio osservare il fenomeno.

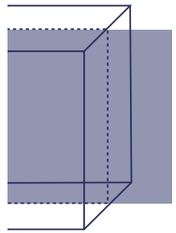
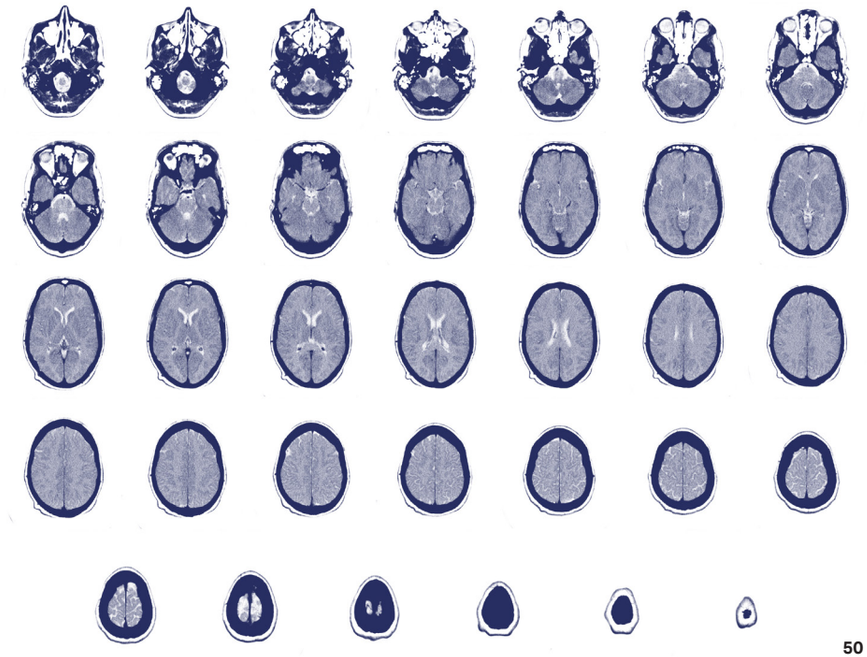
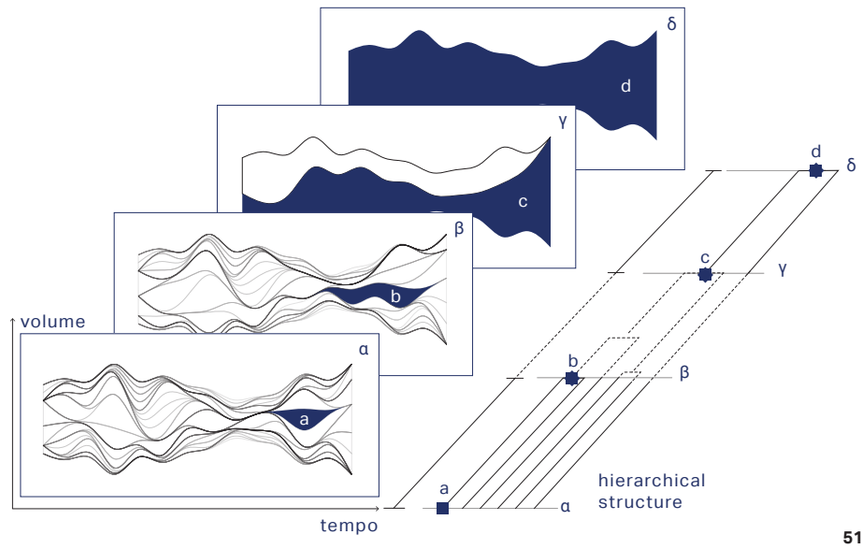


fig.52 Sfogliare uno *streamgraph*.



50



51

fig.50 Una Tomografia Computerizzata (TAC) del cervello umano.
fig.51 Una tomografia dello *streamgraph*.

Tagliare uno streamgraph

Se per visualizzare la profondità del dendrogramma si è sfruttato il concetto di sezione longitudinale, per guardare all'interno dello *streamgraph*, guardarne la composizione (*topic mixture*, *keywords* e relativi documenti) è necessario fare un'operazione di *sezione trasversale*, come se si volesse affettare un solido tridimensionale. Come in un tessuto muscolare così anche uno *streamgraph* si compone di filamenti. Altrettanto nota è l'immagine di un tessuto muscolare affettato trasversalmente. (fig. ☞ 55 e 56)
Quella che, volgarmente, potremmo definire come una fetta di *streamgraph* è dunque una sezione ideale che mostra la configurazione interna di un momento specifico del fenomeno. (fig. ☞ 54)

Per Samuel Ronnqvist et al. in *Interactive Visual Exploration of Topic Models using Graphs*

> *the force directed graph visualization provides a view that more intuitively communicates topic similarity structure*
>
> — *S. Rönqvist et al., 2014*

così, anche nel caso di *TopTom*, la forza che attrae i topic è la forza generata dalla matrice di similarità.
A differenza di *Interactive Visual Exploration of* Topic *Models using Graphs* (S. Rönqvist et al., 2014), in *TopTom* la scelta di rappresentare topic e *keywords* come *metaballs*[64] in continuo movimento suggerisce l'idea di un sistema entropico con transizioni naturali ed armoniche come nel caso dell'efficace visualizzazione interattiva *How riot rumours spread on Twitter*[65] (R. Procter et al., 2011) e il progetto *WorldPotus*[66] (Accurat, 2016) dove il modello visivo a *metaball* suggerisce l'idea di un sistema dinamico.
Applicando l'idea di sezione trasversale *streamgraph*

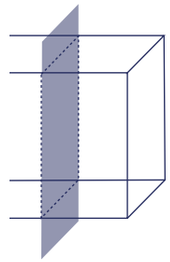
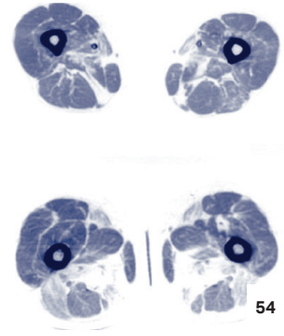


fig.53 Tagliare uno *streamgraph*.

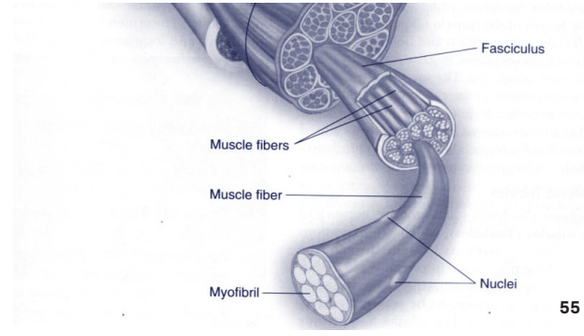
64. Rappresentazione visiva del processo di distaccamento fluido di due elementi circolari.

65. https://www.theguardian.com/uk/interactive/2011/dec/07/london-riots-Twitter

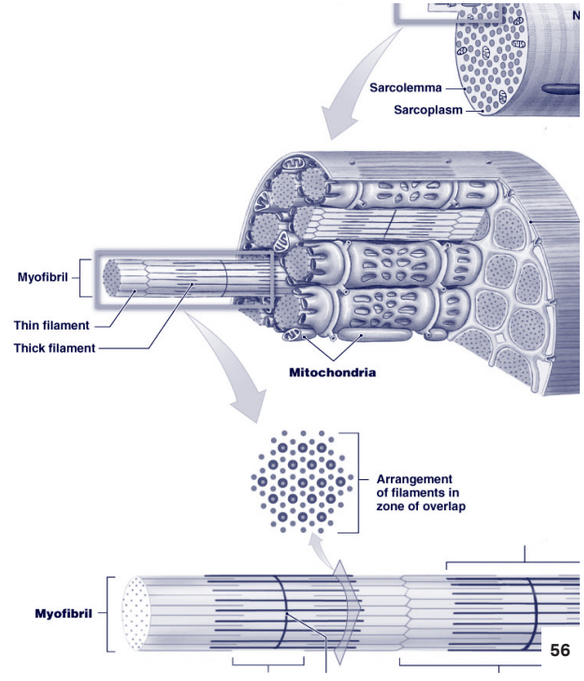
66. http://www.worldpotus.com/#/



54



55



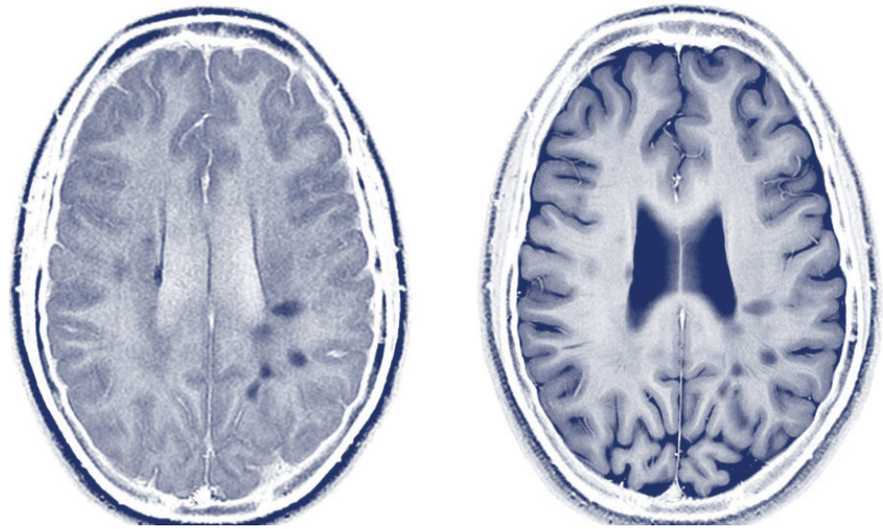
56

fig.54 Una Tomografia Computerizzata (TAC) di due gambe. I muscoli sono in evidenza.
fig.55 Rappresentazione scientifica didascalica della fibra muscolare
fig.56 Rappresentazione scientifica didascalica delle miofibrille.

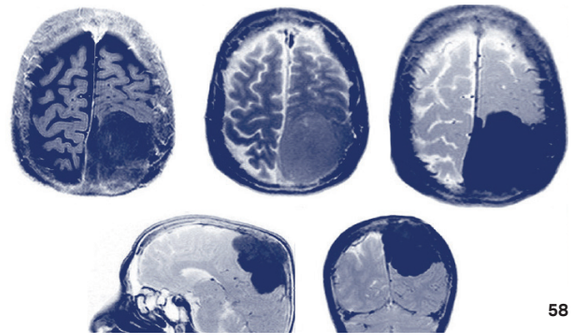
e *force layout* non sono elementi comunicanti in una dashboard come in *TextFlow* (W. Cui et al., 2011) ma condividono la stessa metafora aiutando l'utente ad interagire con lo strumento e a seguire una narrazione.

Anomalie come liquido di contrasto

Negli anni, la rappresentazione dei topic emergenti o anomali si è affidata a variazioni cromatiche come nel caso di *FluxFlow* (J. Zhao et al., 2014) pensato per rivelare comportamenti anomali su Twitter, oppure con la sovrapposizione di elementi direttamente sulle zone anomale di un grafico a flusso (W. Cui et al, 2011).
Per *TopTom* sono stati fatti diversi tentativi: sia visualizzando le anomalie come elementi aggiuntivi di uno *streamgraph*, valutato troppo rumoroso come modello visivo, sia assegnando diverse tonalità di colore ai topic in base all'intensità dell'anomalia. Per quest'ultimo tentativo era necessario però che tutti i topic fossero cromaticamente identici come nel caso di *Emoto* Topic *Explorer* (M. Stefaner, 2012) dove lo *streamgraph* di partenza è monocromatico e il colore indica intensificazione di attività per il paese selezionato.
Tuttavia, la necessità di combinare due diversi tipi di anomalia (*volume increase* e *volume decrease* ☞ 5.1.5) non ha reso possibile questo tipo di scelta, per cui la rappresentazione delle anomalie è stata ripensata in ottica di visualizzazione filtrata, immaginando di sovrapporre allo *streamgraph* policromatico un livello in scala di grigi con zone dal rosso al rosa in base all'intensità della anomalia come se si osservasse una tac effettuata con liquido di contrasto. (fig. ☞ 57 e 58)



57



58

fig.57 Analisi dell'encefalo con liquido di contrasto per delimitare le zone affette da tumore.
fig.58 Analisi dell'encefalo con liquido di contrasto per delimitare le zone affette da tumore.

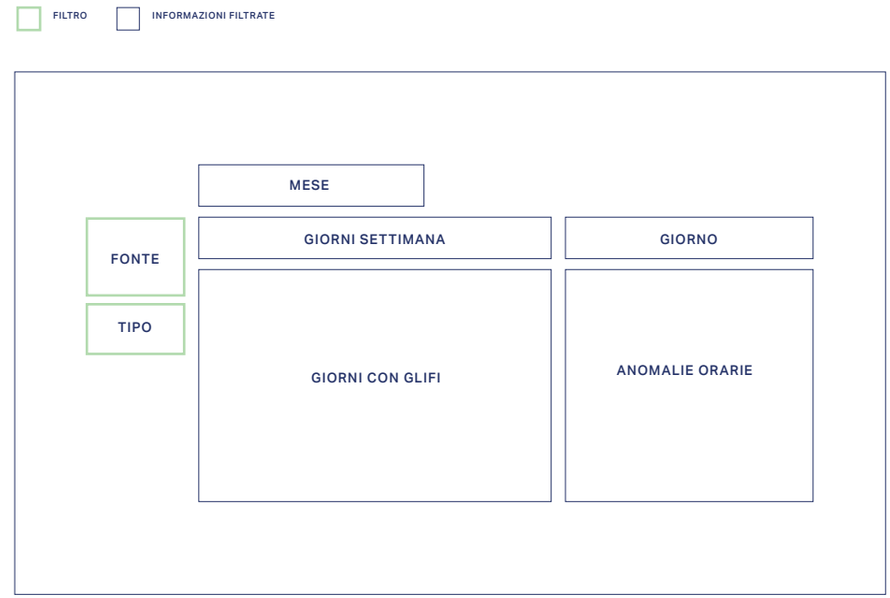
5.4 Lancio, filtri e approfondimento

In an architecture of content, the information becomes the interface.

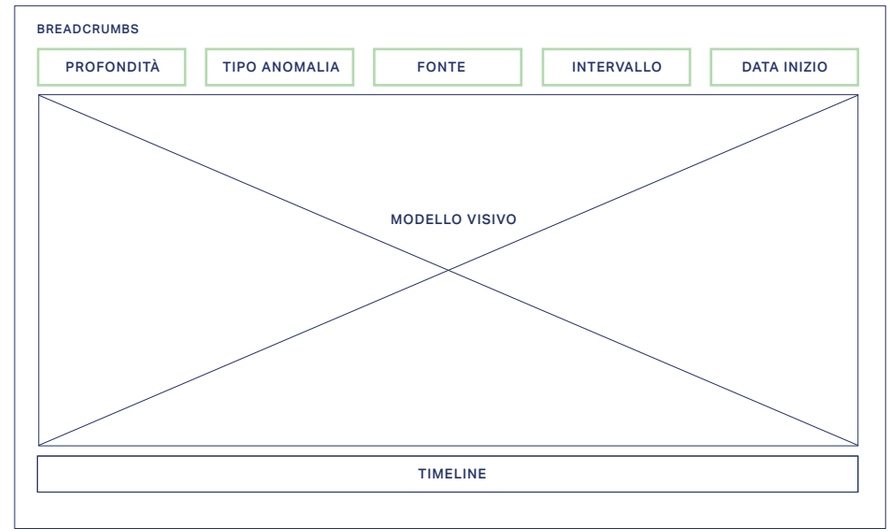
—*E. Tufte, 1997*[67]

67. Tufte, E. R, Visual Explanations: Images and Quantities, Evidence and Narrative, 1997

La progettazione dei *wireframe* è stata un processo lungo che ha richiesto spesso confronto con l'utente. La struttura di base è composta da due modelli di *wireframe*: quella che descrive il calendario (fig. 59) e quella relativa alle varie declinazioni del modello visivo (fig. 60).
Il calendario mostra le anomalie e attraverso i filtri l'utente può personalizzare l'analisi, considerando diverse fonti e diversi contenuti (*keywords* o *entities*).
I filtri più interessanti sono quelli che agiscono direttamente sul modello visivo come la barra della profondità.
In sintesi la struttura generale si può suddividere in tre parti: quella definita dal calendario ovvero la parte di lancio, la parte di approfondimento che consente di esplorare i modelli visivi attraverso approfondimenti e la sovrastruttura dei filtri. Ogni livello di approfondimento può essere filtrato con i filtri della barra di navigazione, il che permette di avere ampi margini di esplorazione. (fig. 61)



59



60

fig.59 Wireframe Calendario
fig.60 Wireframe declinazioni modello visivo.

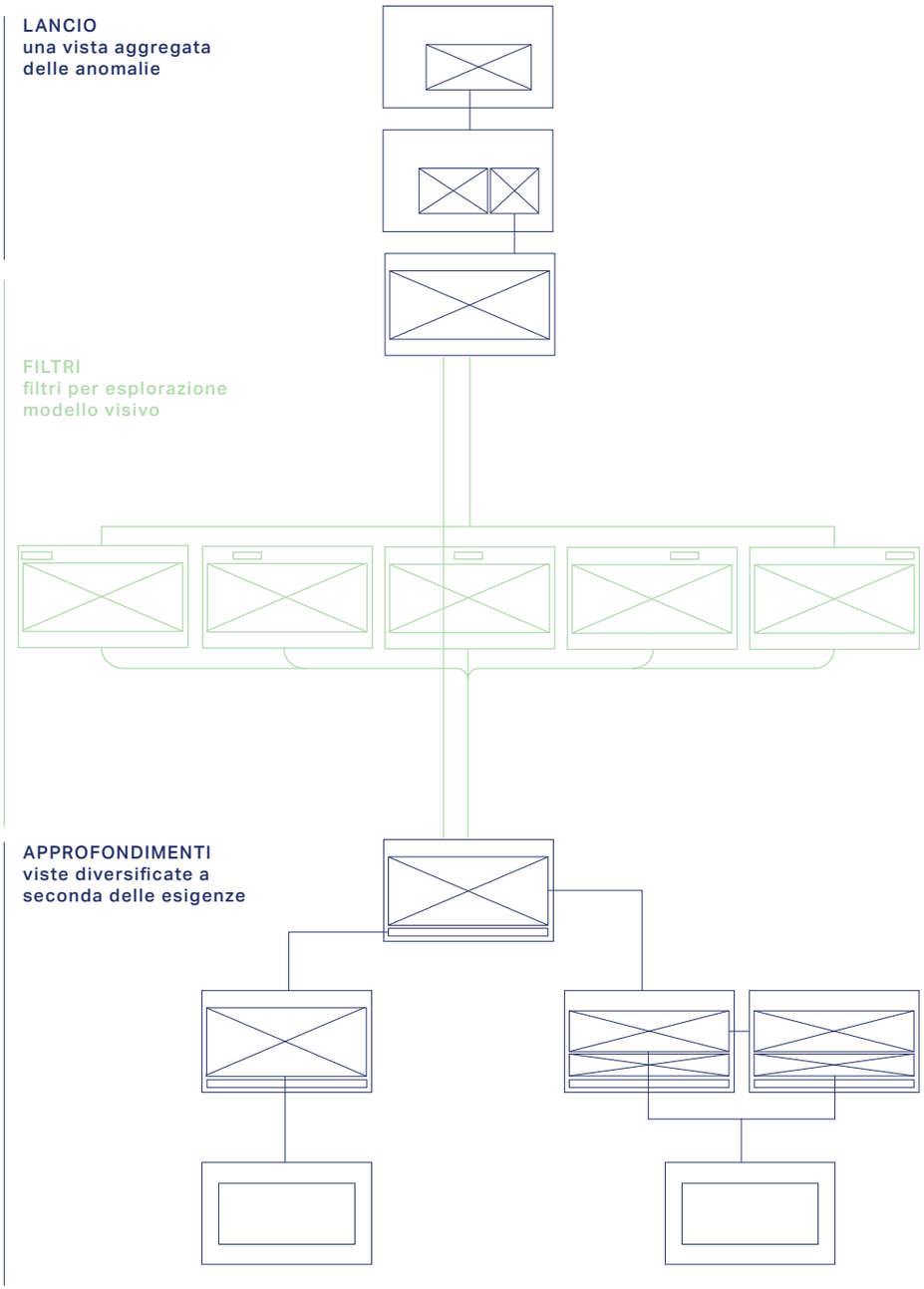


fig.61 Architettura generale di *TopTom*.

Al lancio del tool, la *vista calendario* mostra un'agenda mensile dove su ogni giorno del mese sono indicate le anomalie relative alla giornata; cliccando su un giorno si accede alla vista dettagliata dei topic anomali organizzati per ora.
Selezionando un topic anomalo durante la giornata il sistema va direttamente alla , la *vista focus* che permette di osservare il topic anomalo e i termini principali da cui è composto .
La *vista frequenza parole*, non ancora sviluppata, permette di osservare la frequenza e ripetizione dei termini nel tempo relativamente al topic.
Da qualunque vista contenente la timeline in basso è possibile accedere alla *vista taglio*, che mostra il network di relazioni tra le parole e i topic in uno specifico intervallo temporale, scelto attraverso la selezione di un tassello di timeline.
Infine, dalle viste t*aglio, flusso* e *focus* è possibile accedere direttamente ad un pop con contenuti i link ai documenti caratteristici di quel topic o di quello specifico intervallo temporale. (fig. ☞ 62)

Come in ogni progetto di UI/UX il raggiungimento della versione finale dell'interfaccia è stato un processo lungo e complesso.
Come già mostrato nei capitoli precedenti la progettazione del tool ha richiesto particolare attenzione per permettere una buona sinergia tra interfaccia, modello visivo e dati (sezione ☞ 5.1.5); inoltre una continua modifica dei metadati ha costretto il team ad apportare spesso modifiche all'interfaccia.
La parte più problematica del tool è stata la progettazione della barra di navigazione e della timeline, poiché elementi in grado di gestire e filtrare il modello visivo.

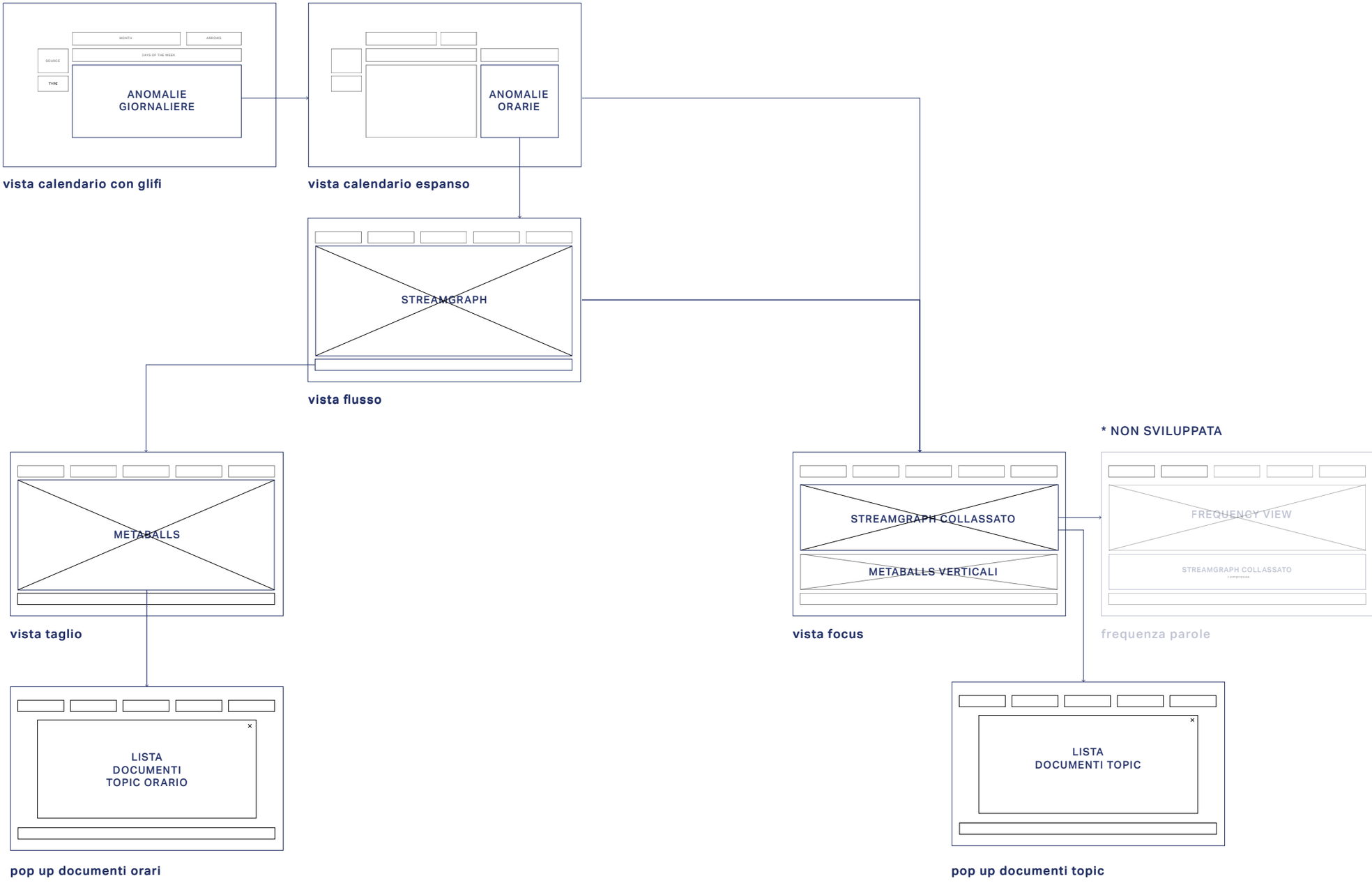


fig.62 Schema di navigazione diretta di *TopTom*.

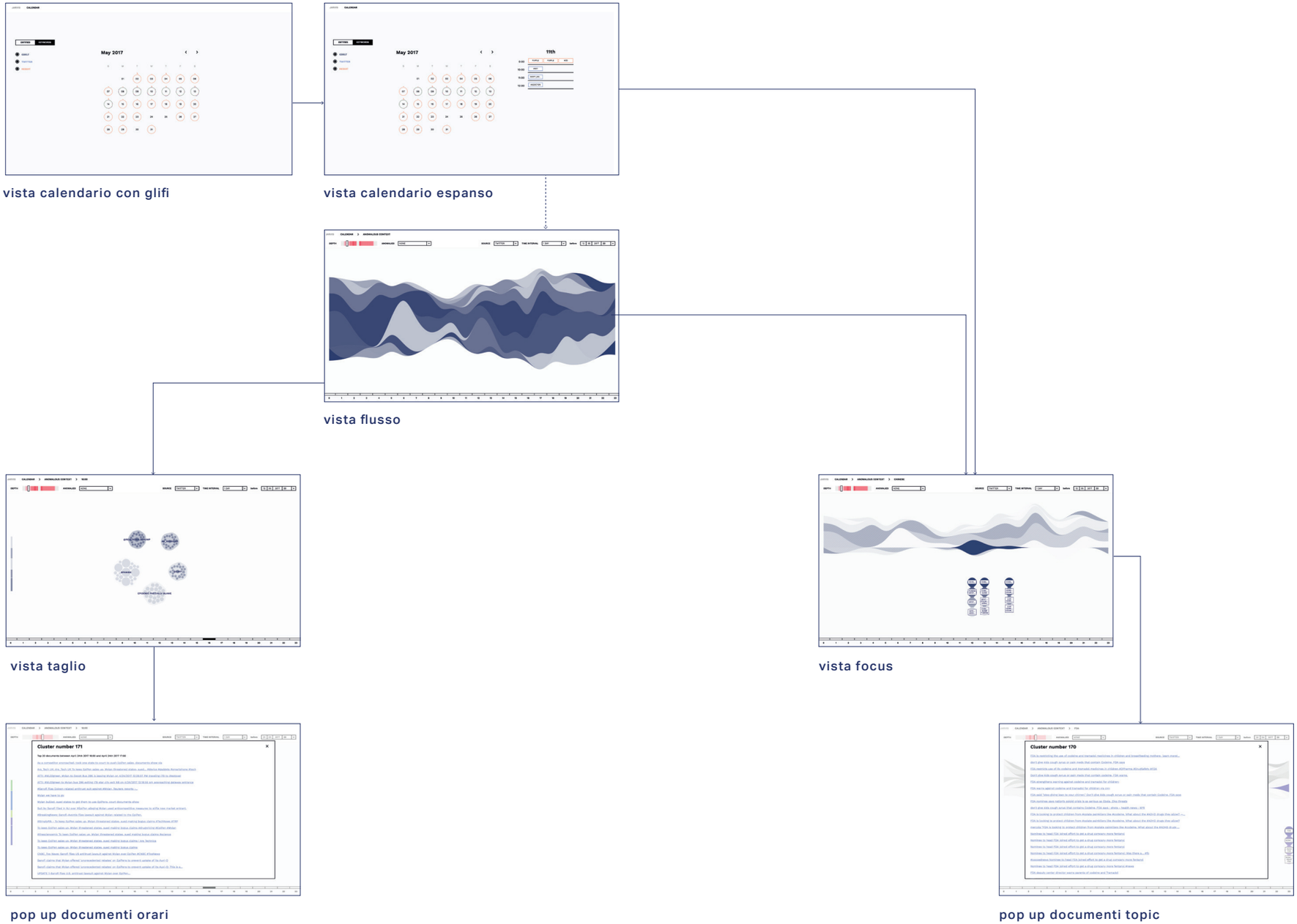


fig.63 Schema di navigazione diretta di *TopTom*.

5.4.1 Applicazione dei modelli visivi

L'interfaccia è strutturata sulla base di quattro viste che includono i rispettivi quattro modelli visivi principali: la *vista calendario* , la *vista flusso*, la *vista focus* e la *vista taglio*; è inoltre previsto una finestra pop up contente la lista dei link esterni ai documenti più rilevanti.
Attraverso la manipolazione diretta del modello diviso l'utente può gestire il grado di approfondimento delle informazioni visualizzate: per esempio dalla vista flusso che rappresenta tutti i topic presenti in un determinato intervallo orario, cliccando su un singolo flusso è possibile isolare l'elemento ed avere informazioni dettagliate sulla composizione interna.
I menù nella parte superiore dell'interfaccia consentono invece all'utente di *filtrare* il modello visivo modificando, ad esempio, il tipo di fonte dei dati.
Inoltre l'interfaccia è pensata per essere navigabile tra le viste utilizzando anche i *breadcrumbs* soprattutto quando è necessario andare a ritroso nei livelli di approfondimento. (fig. ☞ 63)
L'interfaccia di *TopTom* è stata pensata per essere personalizzabile con ogni tipo di dato, quindi con un aspetto neutro. Uno sfondo bianco ottico e una collezione di elementi semplici, definiti da una sottile linea nera sono gli aspetti caratterizzanti del tool, che invece sfrutta una combinazione di colori (generata da *d3scale.js* sulla base di due tonalità di partenza) per differenziare i topic ed indicare le anomalie. (fig. ☞ 64)

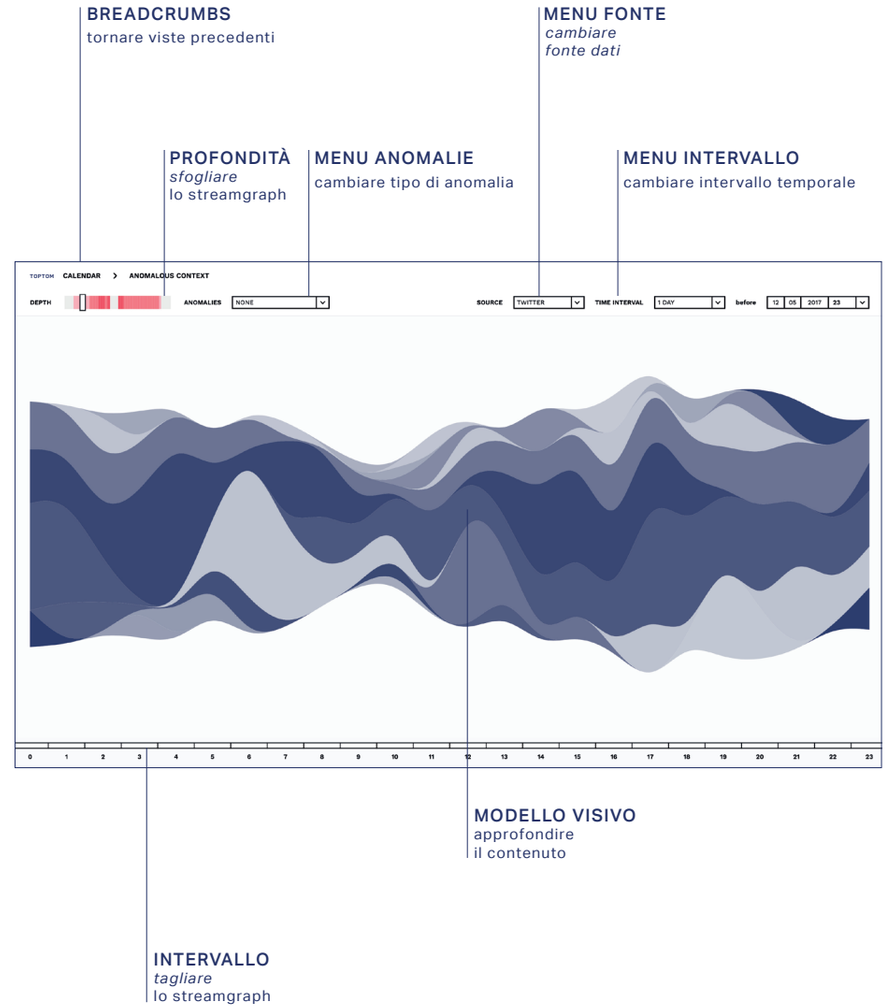


fig.64 Elementi principali dell'interfaccia di *TopTom*.

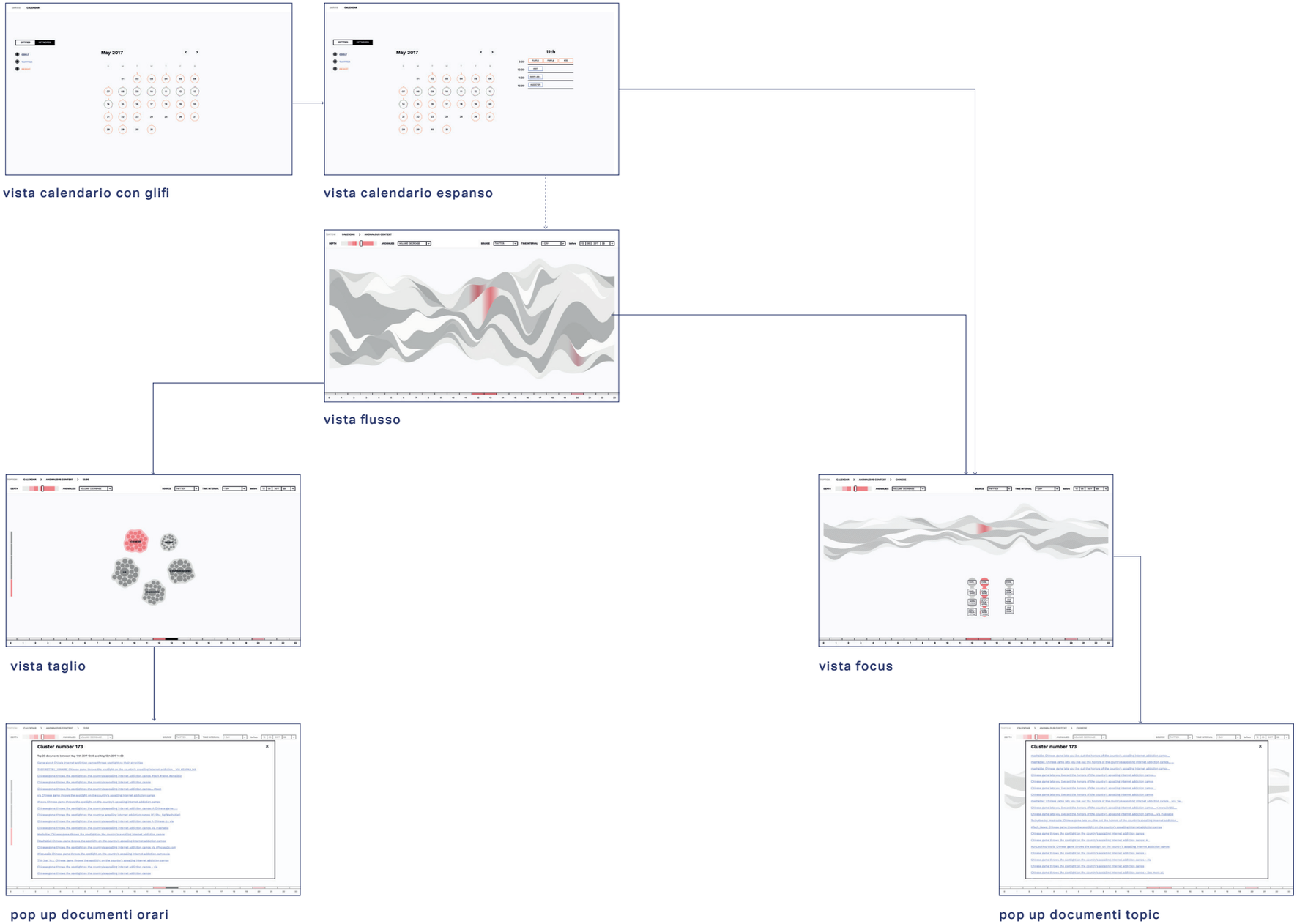


fig.65 *TopTom* con filtro anomalie.

Breadcrumbs

I *breadcrumbs* danno informazioni estremamente approssimative e superficiali rispetto a ciò che può essere visto attraverso il modello visivo.
Tuttavia, permettono di non perdere l'orientamento e, soprattutto, consentono all'utente di poter sempre tornare alla visualizzazione precedente.

Barra di navigazione

La barra di navigazione in alto consente di modificare e filtrare i parametri del modello visivo.

La *depth bar* o barra di profondità, attraverso uno slider, garantisce all'utente la gestione diretta dei livelli di profondità. A differenza dell'interazione usata in *Hierarchical Topics* (W. Dou et al., 2013) che consentiva di osservare i livelli della struttura gerarchica cliccando su un elemento dello *streamgraph*, in *TopTom*, muovendo lo slider da sinistra a destra è possibile osservare come cambi il modello visivo dal minimo al massimo livello di dettaglio. Nel caso di una giornata si può osservare da un topic unico a centoventi topic singoli distribuiti a gruppi di cinque in ogni intervallo orario. Una *heatmap* rossa indica in quali zone gerarchiche sono concentrate la maggior parte delle anomalie.

Il *menu delle anomalie* consente di attivare e disattivare il filtro/liquido di contrasto delle anomalie (*volume increase/decrease*).
La *heatmap* della *depth bar* cambia in relazione al tipo di anomalia selezionato: può essere indicativo rispettivamente di *volume increase* o *decrease* oppure, quando si seleziona none, sommato e sovrapposto.

Il *menu fonte* consente di cambiare il canale online di cui si vogliono analizzare i topic (Reddit, Twitter o GDelt).

La combinazione del *menu intervallo* consente di osservare il fenomeno a diverse granularità temporali.
L'intenzione primaria era quella di progettare una timeline navigabile e interattiva che a causa di un set di dati discontinui non è stato possibile integrare e testare durante il collaudo finale.
Per cui è stato deciso, assieme all'utente, di inserire una tendina che permettesse di inserire l'intervallo temporale desiderato associato alla data di inizio (usando un comune menu a tendina) dalla quale far partire l'analisi a ritroso.

Oltre ad essere un semplice riferimento temporale per la visualizzazione, la *timeline* è anche un elemento interattivo attraverso cui poter navigare da una vista all'altra: dalla vista *streamgraph*, che mostra l'evoluzione dei topic nel tempo alla vista a *metaball* tipica di ogni *intervallo* orario. Per questo motivo è stato necessario progettare le varie configurazioni della timeline sulla base dei possibili intervalli temporali. Se l'utente attiva il filtro delle anomalie, così come la *depth bar* ed ogni altro elemento anche gli intervalli temporali sulla *timeline* si colorano di rosso in base alla presenza di topic anomali. (fig. ☞ 66)

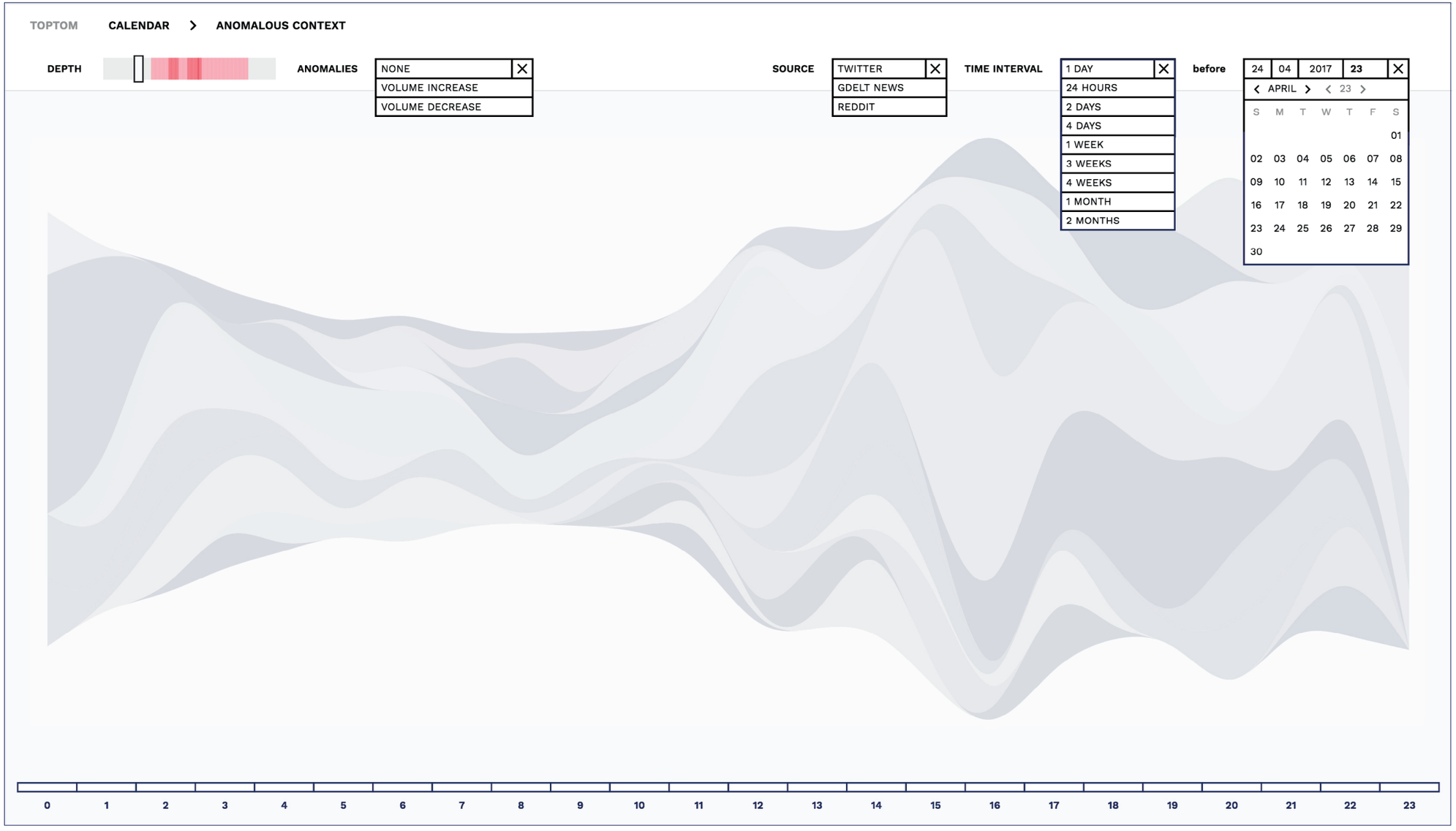


fig.66 Interfaccia di *TopTom* con evidenziata barra di navigazione e breadcrumbs.

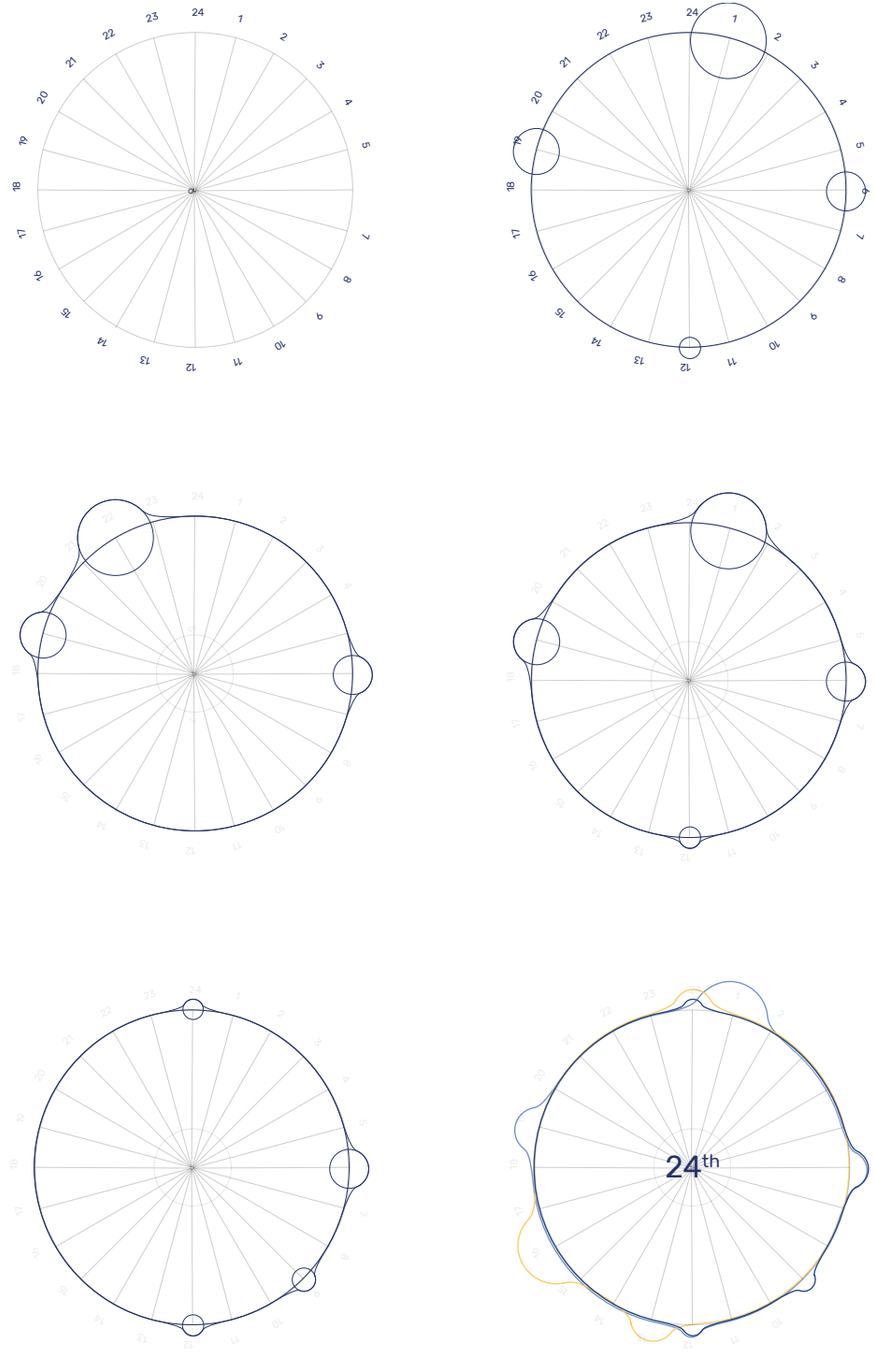
Calendario delle anomalie

Come già mostrato nel paragrafo precedente la vista calendario è il pannello di lancio attraverso cui l'utente può accedere direttamente ad una visualizzazione filtrata rispetto ad una certa giornata, una certa profondità di dettaglio nel dendrogramma, un certo topic e un intervallo temporale di default di 24 ore. Sulla sinistra è possibile filtrare i contenuti del calendario in base alle fonti che si vogliono analizzare (Twitter, Reddit e/o GDelt) ed in base al tipo di analisi (entità o *keywords*). Il calendario è strutturato come un calendario tradizionale che mostra tutti i giorni presenti in un mese e , attraverso le frecce direzionali, è possibile cambiare mese. (fig. ☞ 68 e 69)

Un *glifo*[68] (fig. ☞ 67) a forma circolare è associato ad ogni giornata.
Il glifo riassume l'andamento delle anomalie (se presenti) per quella giornata.
Il glifo è strutturato su base circolare. Una circonferenza, divisa per 24 intervalli, rappresenta una giornata divisa in ore e ad ogni ora un cerchio di diametro proporzionale rispetto al valore rappresenta il volume della anomalia (se presente).

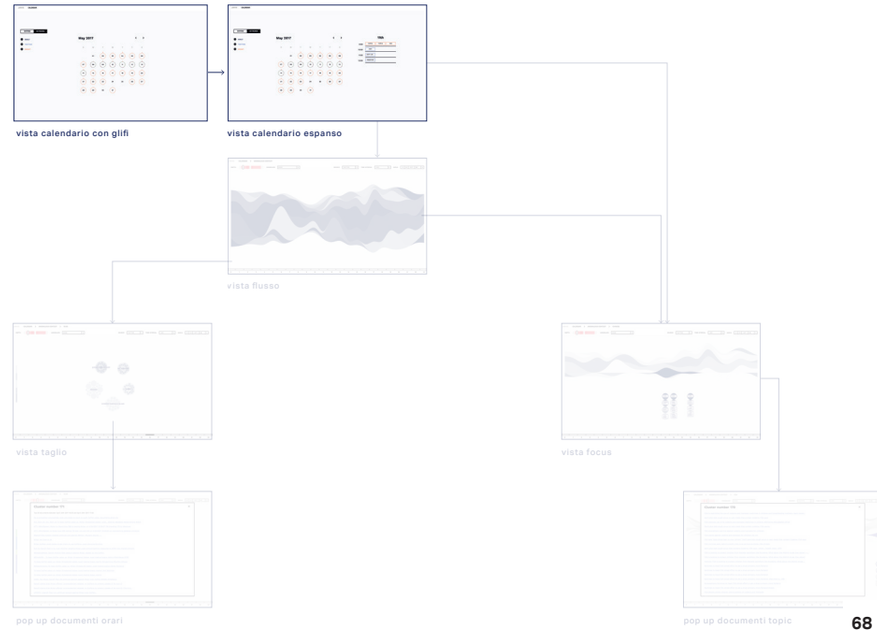
68. Rappresentazione astratta di un grafema, di più grafemi o di parte di un grafema, senza porre attenzione alle caratteristiche stilistiche.

fig.67 Fasi di costruzione geometrica del glifo che circonda le giornate in cui è presente almeno un topic anomalo.



Se il bottone *entities/keywords* è pensato come uno elemento che cambia totalmente i glifi, invece i bottoni delle fonti sono dei *toggles* che consentono di accendere e spegnere le fonti. Infatti, l'utente ha considerato molto utile questo approccio *cross-platform*. Cliccando su una giornata, un nuovo pannello si apre alla destra del calendario e mostra su scala oraria tutte le anomalie presenti nella giornata. Il colore delle etichette, così come della linea di tendenza circolare sul glifo, indica la diversa fonte di dati.
Scegliendo l'azzurro per Twitter e l'arancione per Reddit si è cercato di mantenere i colori tipici delle piattaforme ai fini di aiutare la decodifica da parte dell'utente. Cliccando su una delle etichette che identifica un topic anomalo attraverso la sua parola principale si accede direttamente alla vista filtrata, anomala con le *keywords* espanse.
Dal calendario si accede ad una visualizzazione filtrata che accoglie l'utente nella *vista focus* attraverso la quale è possibile effettuare la maggior parte delle analisi sui topic.

fig.68 Posizione della vista calendario rispetto all'architettura del sistema.
fig.69 La vista calendario mensile e giornaliera permettono di accedere all'elenco dei topic anomali per la giornata scelta.



68

calendario

calendario espanso

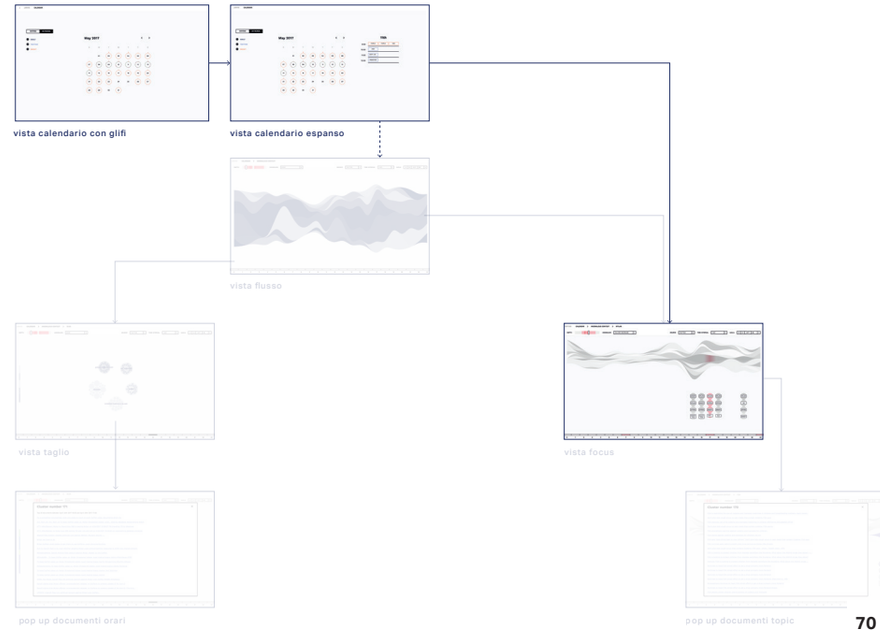
69

Vista flusso

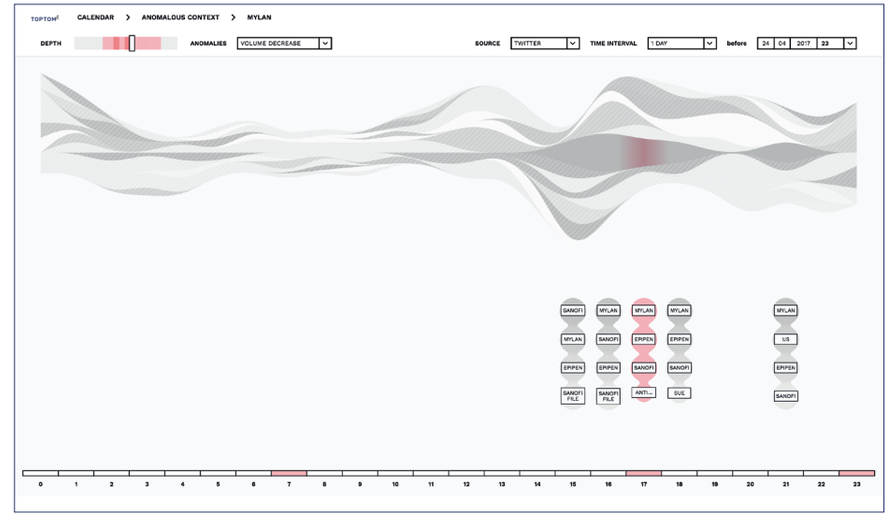
La vista principale che mostra l'andamento dei topic e che è possibile filtrare attraverso la barra di navigazione è uno *streamgraph con layout silhouette ed interpolazione sankey*.

Il *layout silhouette* avvicina i flussi rendendoli omogenei nella forma e permette di osservare a livello macroscopico l'andamento di tutti i flussi/topic e a livello microscopico il comportamento dei singoli topic. L'*interpolazione sankey* gestisce le curve in modo tale che ogni valore zero del volume sia effettivamente rappresentato come zero e non come una media dei valori contigui.

L'utente può manipolare la visualizzazione attraverso hover e click.



70



71

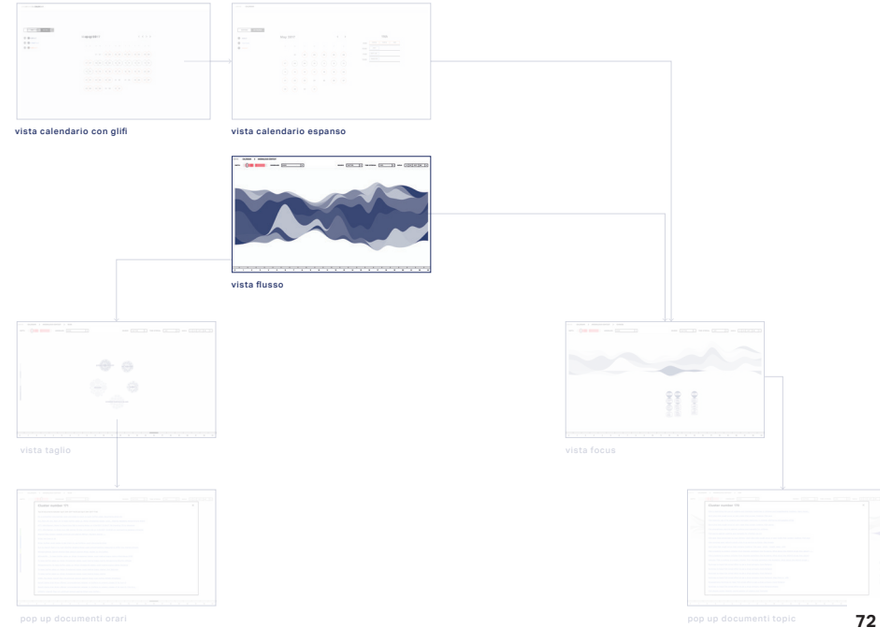
fig.70 Dalla selezione di una anomalia dal calendario si accede direttamente alla *vista focus* con il filtro delle anomalie acceso.
fig.71 *La vista focus* con il filtro delle anomalie acceso mette in luce il topic anomalo che l'analista vuole esplorare.

Primo livello di focus: Hover

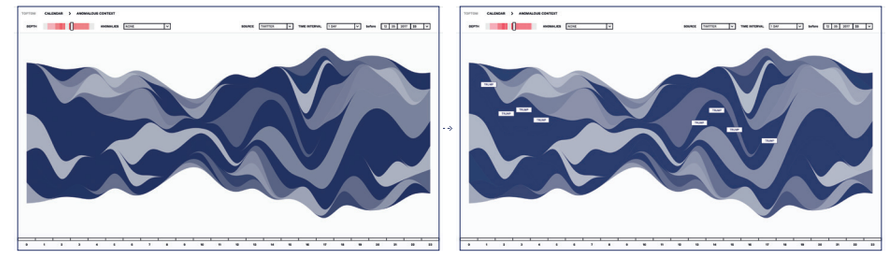
Facendo rollover su ogni singolo stream, un'etichetta mostra la parola più rilevante del topic in corrispondenza di ogni intervallo temporale in cui il flusso è presente.
Questo tipo di interazione è ben conosciuta e usata, come nel caso di *Textflow* (W. Cui et al., 2011) in cui

> *Hovering on a visual element provides users with simplified information, so that they can decide to select it to see more information, or move to other visual elements. For example, when hovering on a* topic, *most representative keywords, which are chosen by our data model, will be overlaid on top of the* topic *layer.*
>
> — *W. Cui et al., 2011*

fig.72 Posizione della *vista flusso* rispetto all'architettura del sistema.
fig.73 *La vista flusso* senza interazioni dell'utente e la *vista flusso* all'hover. Con l'hover l'utente scopre la parola più rilevante del topic per ogni intervallo temporale.



72



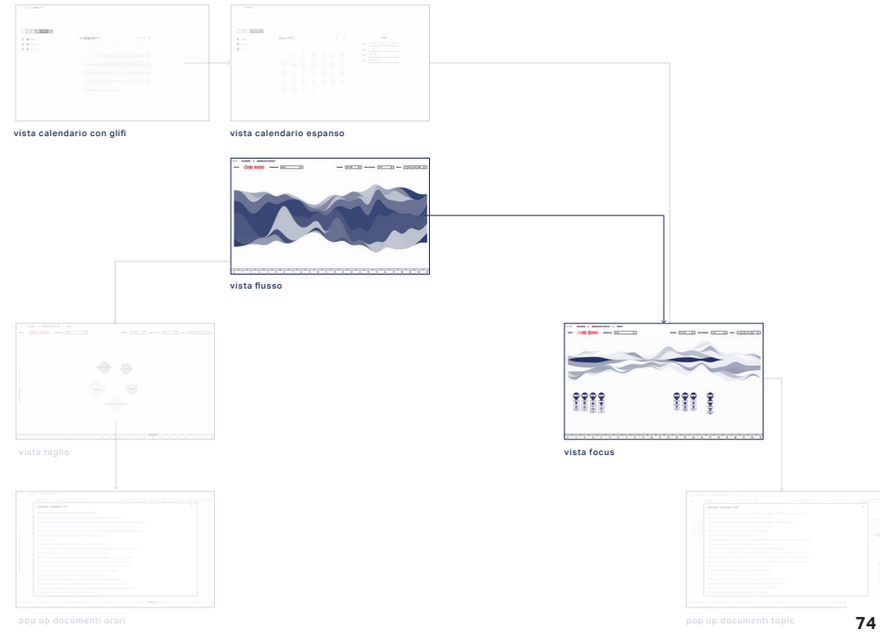
vista flusso

vista flusso / hover su topic

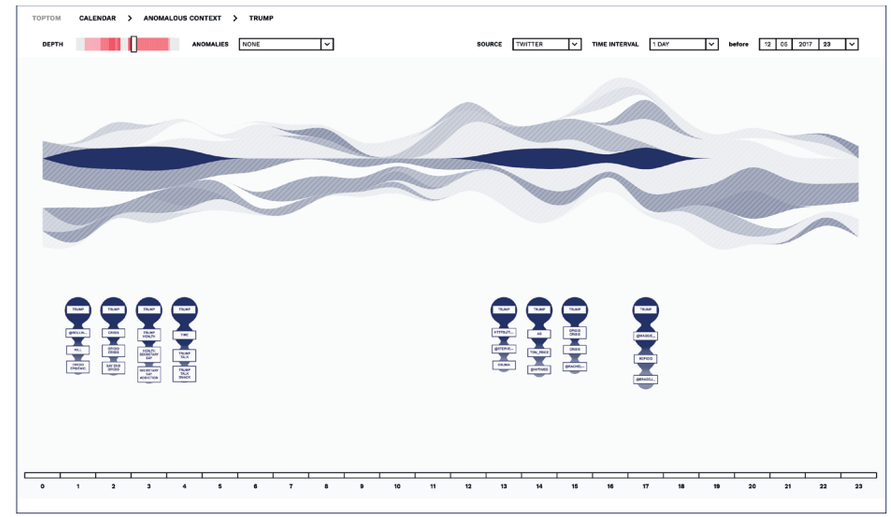
73

Secondo livello di focus: click

La selezione di un elemento consente all'utente di ottenere ulteriori informazioni sull'area selezionata. Una re disposizione collassata dei flussi lascia spazio a delle stringhe di *metaballs* che rappresentano le quattro *keywords* più rilevanti del topic durante gli intervalli temporali in cui esiste.
A differenza di *Textflow*, *TopTom* sfrutta il click solo per scendere in profondità e non per annotare informazioni.



74

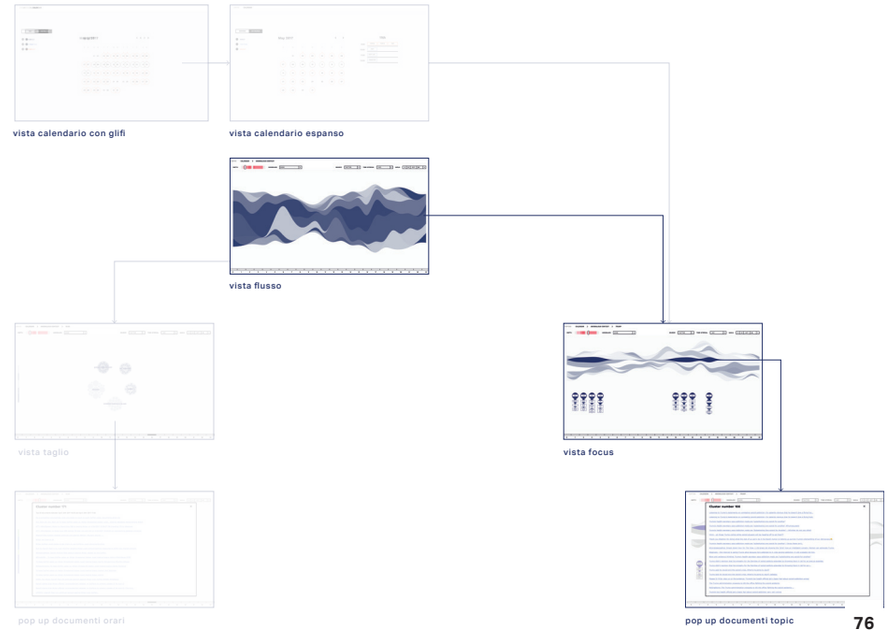


75

fig.74 Posizione della *vista focus* rispetto all'architettura del sistema.
fig.75 *La vista focus* permette di isolare il flusso selezionato, allineandolo al centro, e mostra le quattro parole più rilevanti per l'intervallo temporale in cui il topic esiste.

Terzo livello di focus: click su stream collassato

Cliccando sullo stream isolato, allineato al centro per enfatizzare i volumi, è possibile accedere ai documenti tipici di quest'ultimo: una lista pop up di venti link che indirizzano direttamente al testo originale.
Il colore del link ne indica lo stato (attivo, visitato) e l'ordine indica la rilevanza di ogni documento all'interno di quel topic.



76

77

fig.76 Posizione della *vista documenti per* topic rispetto all'architettura del sistema.
fig.77 La *vista documenti per* topic permette di accedere direttamente ai link agli articoli che caratterizzano il topic selezionato in precedenza lungo tutto l'intervallo temporale.

Vista taglio

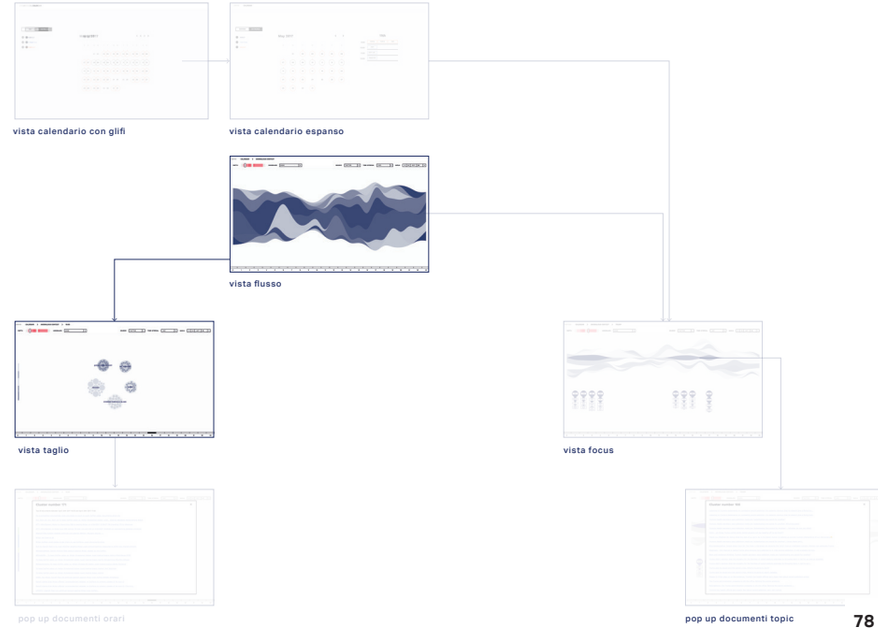
Il secondo modello visivo principale è quello originato dalla sezione trasversale dello *streamgraph*. Ogni ora è caratterizzata da massimo cinque topic (al massimo livello di dettaglio ogni ora avrà cinque topic), ed ogni topic è definito da venti *keywords*, pesate in base alla loro rilevanza.

Selezionando un intervallo temporale nella timeline è possibile osservare al massimo cinque topic presenti alla stessa e la loro composizione in fatto di termini.

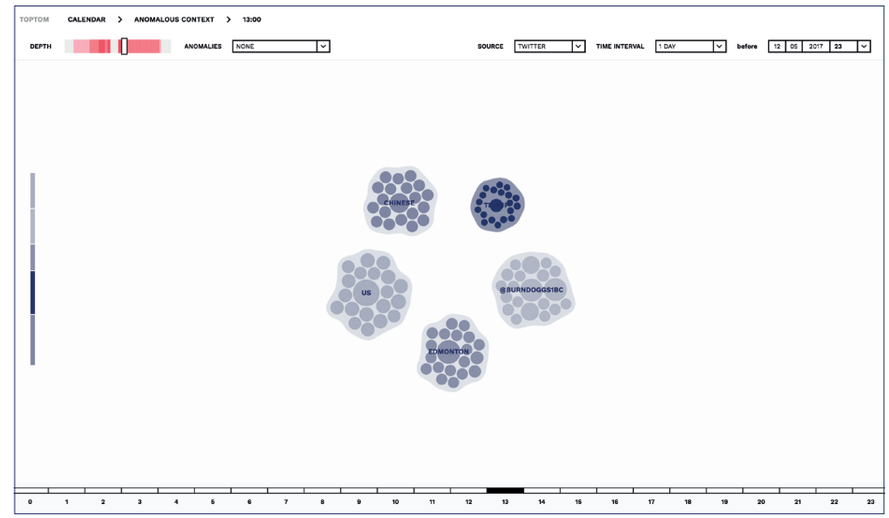
La vista a *force layout/metaballs* comprende un massimo di cinque *metaball* (topic) con all'interno venti *keywords* e la distanza tra i topic è dettata dalla matrice di similarità.

L'area colorata attorno alle *metaballs* consente di dare una forma al topic e confrontarlo con gli altri ad esso contemporanei.

La disposizione delle *keywords* è casuale. Una *barchart* sulla sinistra riassume la proporzione dei topic, come se fosse la parte sezionata dello *streamgraph*.



78



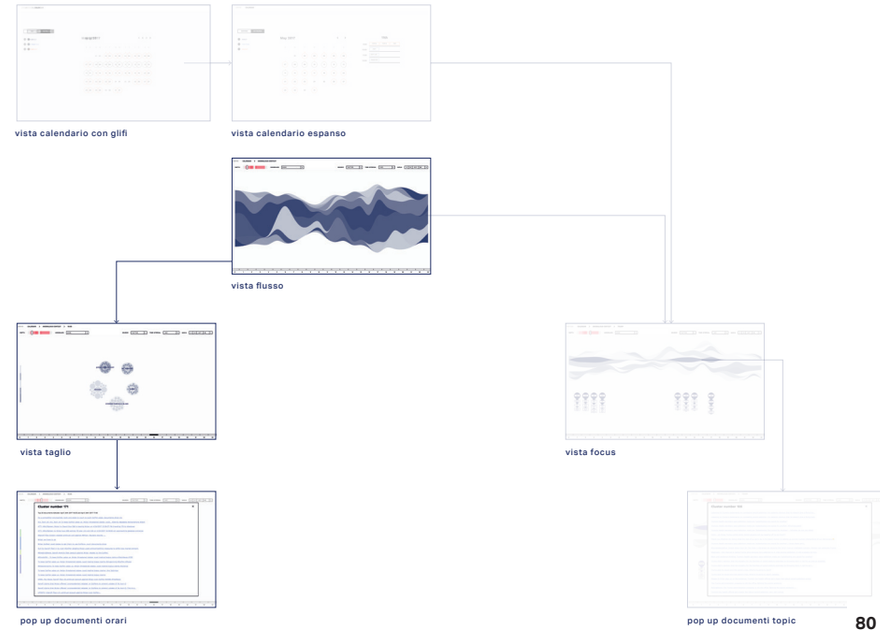
79

fig.78 Posizione della *vista taglio* rispetto all'architettura del sistema.

fig.79 La *vista taglio* permette di accedere direttamente ai link agli articoli che caratterizzano il topic selezionato in precedenza lungo tutto l'intervallo temporale.

Livello di focus: Click su topic

Cliccando su ogni *metaball* è possibile accedere ai documenti tipici di quel topic in uno specifico intervallo temporale specificato anche all'interno della finestra pop up. Per chiudere la finestra è sufficiente cliccare la crocetta in alto a destra, e continuare con l'esplorazione.



80

81

fig.80 Posizione della *vista documenti orari* rispetto all'architettura del sistema.
fig.81 La *vista documenti orari* permette di accedere direttamente ai link agli articoli che caratterizzano l'intervallo temporale specifico(selezionato dalla *vista taglio)* del topic selezionato.

5.5 Design delle URLs e delle API

Nel corso della fase di progettazione la struttura del dataset è stata continuamente soggetta a cambiamenti grazie a un dialogo dinamico tra le componenti (*data science*, design e programmazione) prima di ottenere la forma definitiva che solo a progetto inoltrato è stata tradotta in API.
Progettare un tool che funzioni su chiamate API necessita di un sistema fortemente strutturato lato programmazione.
Una API (Application Programming Interface) è definita da struttura, parametri e risposta.
In *TopTom* il sistema API funziona con due diversi tipi di Api relativi alle viste principali: vista calendario e vista *streamgraph*.
Migliorare la struttura del JSON ha significato rendere più rapida la traduzione delle API in dati ottimizzando le prestazioni del tool.
Inoltre, ai fini di rendere semplice la condivisioni di stati precisi dell'applicazione *TopTom* è stato integrato un sistema di URLs sulla base della struttura delle API molto dettagliato.

fig.82 URLs per gli stati possibili del calendario
fig.83 URLs per i possibili stati delle sezioni dell'interfaccia con i modelli visivi a *streamgraph* e *metaball*.

-	STATO	VALORI	DESCRIZIONE
toptom.com/calendar.html?	date	yyyy-mm	Mese mostrato all'apertura del calendario
	&day	yyyy-mm-dd	Giorno mostrato sul calendario
	&type	[entities/keywords]	Tipologia di dato
	&source	[twitter/reddit/gdelt]	Fonte dati

80

-	STATO	VALORI	DESCRIZIONE
toptom.com/index.html?	date	yyyy-mm-dd-hh-mm	Data impostata dalla barra di navigazione
	&timespan	[24h/1d/2d/3d/4d/5d/1w/2w/4w/1m/2m]	Span temporale d=giorni, w=weeks, m=months
	&depth	[0-119]	Profondità del dendrogramma, il numero corrisponde all'ID del livello
	&anomaly	[none/type1/type2/type3]	Tipo di anomalia
	&slice	[interval/cluster]	Selezione vista stream o circle packing
	&interval	[8h/24h/7d/30d]	intervallo temporale
	&cluster	[0-238]	Numero del cluster, l'ID corrisponde all'ID dei *nodes*
	&clustertype	[focus/frequency]	Tipologia di azione sull'elemento cluster. focus: primo click frequency:visualizzazione fili
	&showdocs	[true/false]	Pop up dettaglio documenti
	&docscluster	[0-238]	Documenti relativi all'ID del cluster. L'ID corrisponde all'ID dei *nodes*
	&docsinterval	[8h/24h/7d/30d]	Documenti relativi all'intervallo scelto

81

5.6 Lo streamgraph è il cuore del sistema!

Il 23 Novembre 2017 si è svolto il collaudo del tool dall'esito positivo, durante le discussioni con l'utente sono emersi alcuni spunti di riflessione finale.

Sia gli analisti che il manager sono stati positivamente impressionati dal design generale, ed hanno definito la visualizzazione innovativa, esteticamente piacevole e con una *user experience* intuitiva.

> *TopTom is an advanced visual tool so that the analyst can both follow the* topic *evolution over time and drill down towards pointwise facts.*
>
> — *SuperUtente*

Quando è stato chiesto agli utenti, a loro avviso, quale fosse la vista più efficiente, tutti si sono trovati in accordo sostenendo che lo *streamgraph* fosse il cuore del tool.

> *the streamgraph is the core of the system.*
>
> — *SuperUtente*

Gli utenti tester continuano sostenendo che la maggior parte dell'attività consiste nell'esplorazione delle diverse granularità temporali, e, una volta selezionata la temporalità più conforme al tipo di analisi, l'esplorazione tomografica dei livelli di aggregazione permette di trovare le informazioni più interessanti e utili.
Alcuni affermano che il taglio temporale è prevalentemente usato per avvicinarsi agli eventi ed osservare le compresenze temporali di diversi topic ma, ovviamente, non è particolarmente utile per visualizzare il comportamento delle anomalie.
Per quanto riguarda le anomalie è stata genericamente apprezzata la possibilità di accendere/spegnere il livello di contrasto.

> *the contrast approach to highlight anomalies that can be activated on every view is very much appreciated*
>
> — *SuperUtente*

5.7 Metafore coordinate

In generale, l'aspetto innovativo del tool risiede in una progettazione coordinata di interfaccia e modello visivo che lascia spazio ad altre innovazioni identificabili nella metafora di partenza, nell'aspetto e nelle funzionalità del modello visivo.

La metafora della tomografia computazionale ha aiutato l'utente a visualizzare tridimensionalmente la struttura dei flussi di topic estratti, sia gerarchica che di composizione interna.
L'anomalia pensata come liquido di contrasto e livello on/off ha permesso di visualizzare le perturbazioni in tutta l'interfaccia senza pesare visivamente sulla struttura del modello visivo.

> *The contrast approach to highlight anomalies that can be activated on every view is very much appreciated.*
>
> — *SuperUtente*

La presenza di un calendario delle anomalie, molto apprezzato dall'utente, ha permesso di riassumere in un modello visivo noto le anomalie caratteristiche del mese, indirizzando l'analista direttamente ad una vista dettagliata con parametri pre-calcolati.

5.8 Mancanze

Collaudo e valutazioni personali in itinere hanno portato a galla alcune mancanze sia dal punto di vista dell'interazione che dal punto di vista delle possibilità di analisi offerte dallo strumento.

Interazione

☞ La vista calendario non offre né la possibilità di vedere anomalie su diverse scale temporali, né la possibilità di avere informazioni se si clicca sul numero del giorno stesso;

☞ L'interazione a tendina che consente di cambiare la data di partenza è intuitiva ma macchinosa. Questo problema potrà essere risolto quando sarà possibile lavorare con un set di dati completo, che consenta di poter progettare una timeline interattiva espandibile e navigabile.

☞ Sarebbe interessante introdurre un elemento che consenta di condividere la vista senza dover copiare e incollare l'URL.

☞ Introdurre un bottone "play" che consenta di vedere come le *metaball* si formano e deformano durante l'intervallo di tempo selezionato.

Analisi

☞ La *vista word frequency* non offre molte possibilità di analisi, e al momento, non è ancora stata programmata per la versione online interattiva. Un test dell'interazione potrà far emergere spunti più dettagliati su eventuali sviluppi futuri;

☞ Sarebbe interessante effettuare analisi di come singole *keywords* passano da un topic all'altro sia a livello temporale che gerarchico.

☞ Infine, un ulteriore aspetto interessante sarebbe poter analizzare la robustezza del topic (sia a livello temporale che gerarchico) al fine di fornire più informazioni all'utente già dall'aspetto iniziale del topic;

☞ È stata espressamente richiesta dall'utente finale la presenza di una barra di ricerca in grado di identificare i topic in cui quella keyword è presente;

> *A free text search in the interface able to query the aggregated results and to jump straight to the* topic *containing the word of interest would help the analyst to include domain expert knowledge in the visual* topic *exploration.*
>
> — *SuperUtente*

Le mancanze identificate nel progetto hanno permesso di delineare un programma di possibili lavori futuri per implementare la piattaforma e migliorarne l'efficienza.

☞ Colmare le lacune emerse sia lato interazione che analisi, partendo dalle prime due mancanze legate all'interazione per poi concludere con le altre. Le mancanze identificate nell'analisi sono colmabili ma richiedono una ricerca più approfondita, che vada ad indagare bibliografia e casi studio.

☞ Poiché *TopTom* nasce come tool di monitoraggio sarebbe interessante integrare uno stream di dati che indichino gli eventi, per avere riferimenti storici rispetto all'elaborazione dati dell'algoritmo.

☞ Un'interazione in tempo reale su un grande pannello di controllo, suggerita anche dall'utente esperto di dominio durante un seduta di revisione e feedback, potrebbe migliorare l'esplorazione e il monitoraggio del fenomeno, sfruttando, ad esempio, una tecnologia touchscreen.

5.9 Verso una nuova sperimentazione

A seguito dello sviluppo della piattaforma, delle analisi delle mancanze e dei lavori futuri, sono emersi due principali *insights:*

☞ I meccanismi alla base della *topic detection* non sono molto semplici da capire per un utente poco esperto.

☞ Il topic in sé è un concetto astratto, privo di un'identità. Esiste solo quando definito da altri parametri che ne sottolineano la complessità.
È a partire dal secondo *insight* che nasce l'ultimo capitolo di questa tesi.

Vidi ‘l maestro di color che sanno seder tra filosofica famiglia.

— Dante Alighieri, Divina Commedia, Inferno, IV

6. De Topic

6.1 Visualizzare algoritmi

Mike Bostock sostiene che gli algoritmi siano un affascinante caso studio per la data visualization, infatti ha dedicato parte delle sue ricerche alla visualizzazione del funzionamento di queste misteriose scatole nere, sfruttando tutte le potenzialità della programmazione front end.

> *Algorithms are also a reminder that visualization is more than a tool for finding patterns in data.*

Visualization leverages the human visual system to augment human intellect: we can use it to better understand these important abstract processes, and perhaps other things, too.

— *M. Bostock, 2014*[69]

Il focus del progetto *TopTom* non è quello di visualizzare il comportamento dell'algoritmo ma le possibili rappresentazioni dei suoi risultati. Tuttavia, alcune caratteristiche del dataset associate al modello visivo con cui i dati sono stati visualizzati hanno fatto emergere considerazioni interessanti in merito alla rappresentazione del topic.

Il topic in sé non esiste, la statistica etichetta il topic con un numero cardinale, in alcuni casi in maniera arbitraria si sceglie di definire un topic tramite la sua keyword principale (*TopTom*) o tramite, l'argomento, appunto, di cui tratta prevalentemente (F. Wei et al., 2010). La scelta è però arbitraria e rischia di lederne la complessità. L'esistenza del topic è legata ad altri parametri: *gerarchia*, *temporalità contenuto* e *relazioni*, che sono i parametri tassonomici elencati precedentemente (cap. ☞ 04).

69. https://bost.ocks.org/mike/algorithms/

6.2 Definire il topic

6.2.1 Un luogo, comune

Nel linguaggio odierno, secondo i siti web dei tre dizionari online più visitati al mondo (*fonte: Alexa.com*[70]) al termine topic è ormai comunemente attribuito come primo risultato il significato di *subject*, *theme*.

Wordreference.com[71]
a subject of discussion
the subject of a speech or piece of writing

Wiktionary.org[72]
Subject; theme; a category or general area of interest
(obsolete) An argument or reason
(obsolete, medicine) An external local application or remedy, such as a plaster, a blister, etc.

Thefreedictionary.com[73]
The subject of a speech, essay, thesis, or discourse.
A subject of discussion or conversation.
A subdivIsion of a theme, thesis, or outline
(Linguistics) A word or phrase in a sentence, usually providing information from previous discourse or shared knowledge, that the rest of the sentence elaborates or comments on. Also called theme.

Una delle definizioni più interessanti ed utili a fini di

70. https://www.alexa.com/
71. http://www.wordreference.com/
72. https://www.wiktionary.org/
73. http://www.thefreedictionary.com/

ricerca viene proposta da *Thefreedictionary.com*, infatti, sfruttando la disciplina della linguistica, allude ad una complessità sottesa al lemma topic definendolo come un elemento all'interno di una frase che fa riferimento ad una conoscenza condivisa, come se il topic fosse un segno per comunicare attraverso una singola espressione tutto ciò che è relativo ad un determinato ambito migliorando la comprensione da parte di tutti i destinatari del discorso. Quest'ultima definizione di topic è quella che più si avvicina alla definizione etimologica del termine.

Infatti, nonostante la sua natura anglofona, il termine topic deriva etimologicamente dal greco antico ἡ τοπικός (*e topikos*) che significa locale, del luogo.
L'utilizzo al plurale τά τοπικα (*ta topika*) del termine, tradotto *I* Topic*i* deve la sua notorietà allo scritto di Aristotele[74], parte dell'*Organon*, famosa raccolta di trattati del pensatore di Stagira riguardo la Logica, utilizzata come strumento didattico nella scuola da lui fondata il Liceo.
Ne *I* Topic*i* Aristotele affronta il tema della dialettica sostenendo che in un dialogo tra due o più individui, una discussione nasca sulla base di dati e informazioni su cui non c'è vera certezza e che si giunga alla conclusione attraverso un dialogo di inferenze probabili.
Infatti, nella rivisitazione del τά τοπικα di Cicerone del 44 a. C la dottrina della *topike* viene definita come *l'arte di trarre argomenti dai luoghi comuni.*
Trarre argomenti da luoghi comuni è ciò che in estrema sintesi fa un algoritmo di *topic modeling*. Se si prova a sostituire il termine *argomenti* con topic e *luoghi comuni* con *parole*/topic *comuni* si ottiene all'incirca la definizione della funzione principale di un algoritmo di questo tipo.

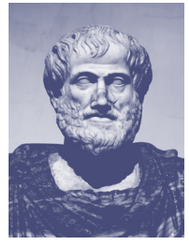


fig.84 Aristotele

74. Stagira 384 a. C – Calcide 322 d. C. Filosofo, scienziato e logico greco antico.

Trarre *topic* da *parole comuni* .
Trarre *topic* da *topic comuni.*

Il dialogo di inferenze probabili di cui parlava Aristotele può essere associato alla definizione di *topic modeling* che fornisce Blei, padre del probabilistic *topic modeling*

> *A topic is a probability distribution over terms*
>
> — *D. Blei, 2003*

L'ambiguità del termine topic è evidente nella moltitudine di sfaccettature di significato: topic è sia *division* che *point*, sia problema che risoluzione, sia *moot point* che *matter in hand.*

6.2.2 Identità del topic

L'identità del topic è complessa e frammentaria per cui è necessario mantenerla tale in ottica di studio e ricerca. La sfida è dare informazioni sul topic senza che risultino informazioni approssimative.
L'identità visiva può essere veicolata tramite immagini, colori, forma o una particolare organizzazione interna degli elementi suddivisa per campi semantici.

Esperimenti preliminari hanno consentito di escludere la definizione dell'identità per campi semantici delle parole che compongono i topic: per fare questo un'ipotesi sarebbe stata affidarsi a *WikiData*[75] ponendo però, per questioni di tempistiche, le basi di una nuova tesi magistrale.
Infatti, l'insieme di algoritmi usati per produrre i risultati mostrati *TopTom* identifica parole *di* o *trigamma*[76] e Wikidata diventa complesso da usare a meno che non si pensi a un protocollo specifico per catalogare correttamente parole chiave formate da più lemmi all'interno di Wikidata.
Era necessario trovare, all'interno dei dati forniti, un elemento che definisse meglio di altri l'identità del topic.

75. Wikidata è un archivio di dati può essere utilizzato da altri, come le wiki della Wikimedia Foundation.

76. Parole composte da due o più lemmi

6.3 Topic, un argomento di conversazione

Quello che nella dialettica aristotelica era affidato all'insieme di conoscenze comuni dei soggetti coinvolti in una discussione, oggi, negli algoritmi di *topic modeling* è affidato al *machine learning*.
Nel caso studio *TopTom* una delle più grandi difficoltà è stata quella di cercare di non ledere l'identità di ogni topic. Infatti, per essere studiato, un elemento va identificato e visualizzato.

Si provi ad immaginare il *corpus* di tweets come un dialogo tra più persone e i topic come argomenti di conversazione in una discussione. La definizione dell'argomento è di per sé un processo complesso. Aristotele ne *I Topici* riferendosi all'ambito dell'argomentazione di un discorso in un dialogo tra due persone, enuncia che

> *[...] Considerando tutte le argomentazioni trattate, si deve inoltre osservare, che a proposito di alcune è più facile cogliere gli elementi sui quali si appoggiano per ingannare l'ascoltatore, mentre a proposito di altre questo è più difficile [...]*
>
> — *Aristotele, I Topici 11*

Analogamente, anche la definizione di un argomento o topic può essere più o meno complessa in base ai contenuti che caratterizzano il topic analizzato.
Sempre ne *I Topici* Aristotele destruttura l'argomen-

tazione ai fini di mostrarne la complessità e la fallacia, poiché l'argomentazione e la dialettica sono le arti che fanno passare per vero soltanto ciò che è probabile.
Non essendo questo un trattato di filosofia, non ci saranno approfondimenti in merito alla struttura dell'opera ma verrà sfruttata analogamente l'operazione di destrutturazione per provare a definire attraverso modelli visivi noti l'identità del topic.
La destrutturazione del topic comincia dalla struttura dei dati che lo definiscono.
Tuttavia, se è relativamente semplice definire il topic come argomento, fonte generale da cui si traggono argomenti, non è così semplice definire quali siano gli elementi che definiscono un topic.

6.4 Destrutturare il topic

Così come nella maggior parte dei casi, in *TopTom* l'identità del topic è definita da un id numerico e da una serie di metadati organizzati in un JSON (cap. ☞4).
Gli elementi contenuti nell'*array* di un nodo descrivono il topic attraverso parametri che possono essere categorizzati secondo le classificazione proposta nel capitolo 4 (*relazione, contenuto, cronologia* e *gerarchia*).

```
"id_node": {
"common_words": [ … ],  (contenuti - relazioni)
"stream_vector": [ … ],  (tempo)
"topic_documents": [...]  (contenuti - relazioni)
"children" : [ ];  (gerarchia - relazioni)
"anomalies":[ ];  (contenuti)
"similarity": [ ];  (gerarchia - relazioni )
}
```

L'id del nodo è il numero con cui viene definito il topic.
Gli elementi che definiscono il nodo si possono dividere per:

☞ Relazioni (parole, documenti,similarità e *children*)

☞ Contenuti (parole contenute nel topic, anomalie, documenti)

☞ Cronologia (momenti dell'arco temporale in cui un topi esiste)

☞ Gerarchia (la struttura gerarchia di figli/genitori)

La definizione di topic, alla latina, come detto precedentemente, ha un'accezione spaziale. La sfida è stata dunque quella di trovare un parametro che potesse definire il topic come un elemento a sé stante, delimitato da confini più o meno definiti in base ad una caratteristica interna del topic.

Le possibili strade di ricerca erano due:

☞ Focalizzarsi sulla rappresentazione visiva di un singolo topic ad una certa profondità di dettaglio e cercare di definire l'identità visiva attraverso la combinazione degli altri parametri (contenuti, tempo, anomalie, gerarchia...);

☞ Concentrarsi su un singolo parametro ed esplorarne le potenzialità.

Era necessario trovare un parametro che potesse raccontare la *storia* di un topic senza lederne la complessità. Durante alcuni incontri con il cliente (cap. ☞5) è emersa la necessità di individuare e rappresentare la robustezza del topic sia dal punto di vista temporale che dal punto di vista gerarchico, ma non era consentito sovraccaricare troppo la visualizzazione *streamgraph* e, inoltre, non era stato possibile definire quali fossero i parametri ideali per definire la robustezza.
Per fare questo era necessario trovare un focus che non fosse direttamente correlato al tempo e al contenuto per non ripetere la visualizzazione a *streamgraph* sperimentata in *TopTom*.

6.4.1 Questioni di famiglia

Il campo *children*, che comunemente viene visualizzato con un dendrogramma, nel tool *TopTom* dà la possibilità di:

☞ Sfogliare lo *streamgraph*

☞ Osservare il dendrogramma a diversi livelli di dettaglio.

Il dendrogramma è la rappresentazione più comune che viene utilizzata da statistici e data scientists per rappresentare le gerarchie. Nel caso di *TopTom* l'algoritmo di clusterizzazione che clusterizza i topic in maniera gerarchica è UPGMA ma esistono molti altri tipi di algoritmo che permettono di ottenere simili risultati (W. Cui et al., 2014) .
Nell'interfaccia di *TopTom* un'attività di analisi mancante è quella di confronto tra topic sulla base dell'algoritmo di clusterizzazione, per cui è sembrato interessante riflettere sull'identità del topic partendo dal concetto di profondità per provare empiricamente a confermare che l'identità dei topic può essere definita dalla robustezza.

La domanda iniziale è stata comprendere se esistessero topic più stabili di altri e da cosa dipendesse la loro stabilità, e inoltre se fosse possibile trovare un modello visivo in grado di dare informazioni sull'identità del topic senza partire dai dati relativi ai contenuti dei documenti, partendo dunque dall'ultima fase dell'algoritmo per cercare di comprendere il contenuto dei topic. Un'osservazione del dendrogramma consente di identificare visivamente topic *deboli* e topic *forti* (fig. ☞ 85), ma non è intuitivo trovare analogie di

LEGENDA

Topic

Cluster

Livello
di aggregazione

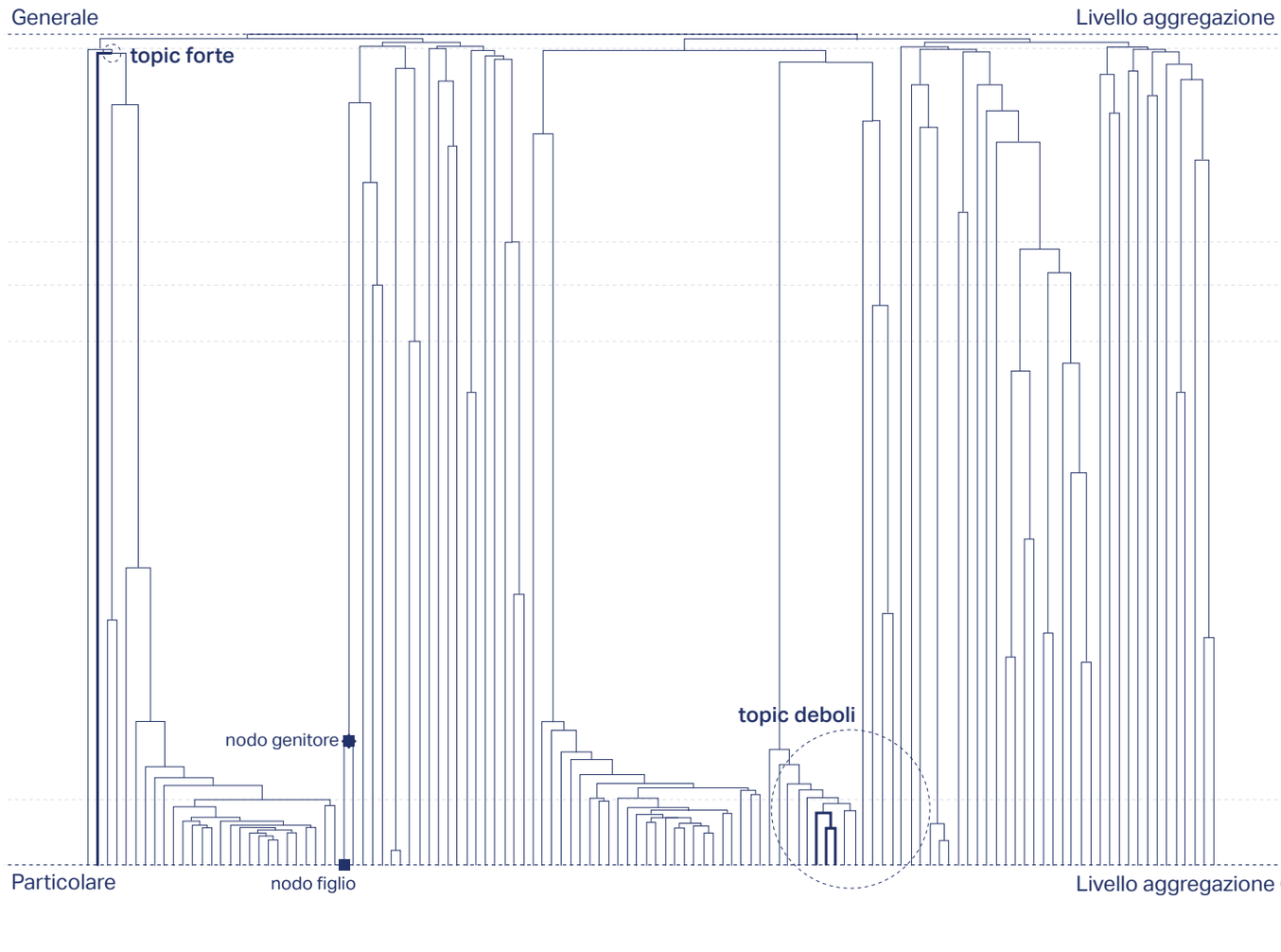


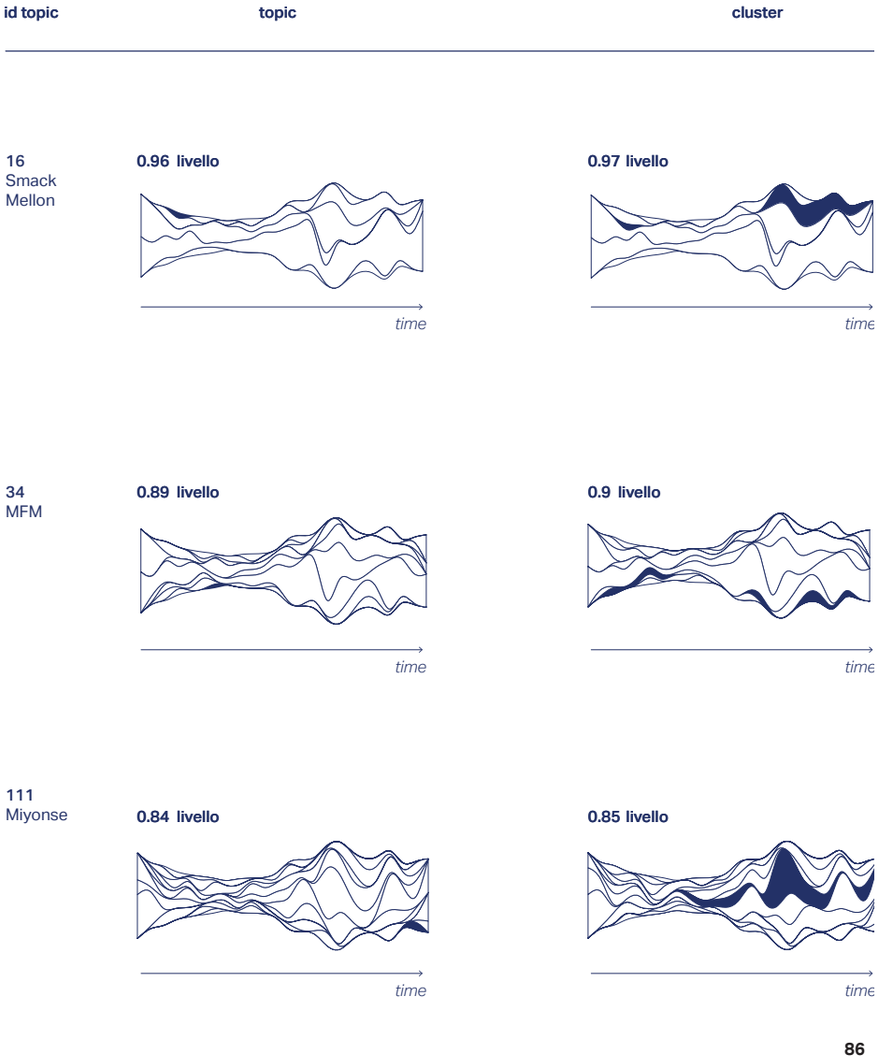
fig.85 Il dendrogramma, modello visivo usato per la rappresentazione di strutture gerarchiche con evidenziati gli esempi di topic deboli e topic forti.

comportamento tra topic ed osservare lo sviluppo dei topic singolarmente, dalla forma più dettagliata a quella più generica.
Nel caso di un set di dati di una giornata composta da 24 ore due topic possono associarsi per similarità ad un livello compreso tra 0 e 1. Il grado di similarità tra due topic è calcolato in base alla similarità tra *keywords* che compongono i documenti contenuti all'interno del topic. Due topic molto simili in fatto di *keywords* saranno probabilmente clusterizzati ad un livello molto basso del dendrogramma.

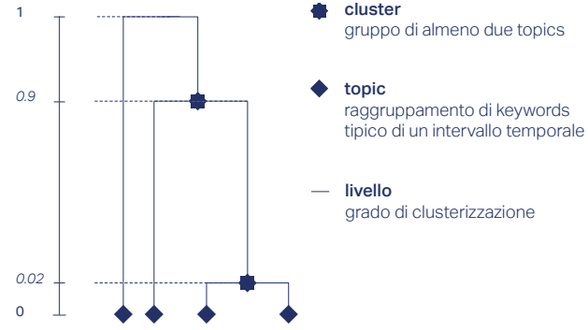
6.4.2 Topic forti e topic deboli

Prima di parlare di stabilità è necessario fare una riflessione sulla distinzione tra topic deboli e topic forti.
Come in una conversazione tra tre persone che parlano del trasporto pubblico a Milano due persone parlano dello sciopero dell'otto marzo e una persona parla della legge sindacale che definisce il diritto allo sciopero, è chiaro come, osservando la discussione ad un livello più generico di dettaglio le due persone che parlano dello sciopero dell'otto marzo stiano parlando di un tema molto generico e quindi debole, mentre la persona che sta parlando della legge sindacale sta trattando un argomento molto specifico, quindi più forte.
Per cui, esistono argomenti deboli che devono unirsi e fondersi ad altri per acquistare forza ed esistono argomenti più forti che mantengono le loro caratteristiche a diversi livelli di profondità di dettaglio, come nel caso della legge sindacale.

I topic singoli (*foglie*) dall'apparente identità debole sembrano quelli che si clusterizzano per primi in prossimità della base del dendrogramma, mentre i topic più robusti sono quelli che si accorpano in prossimità della parte più alta del dendrogramma, che nel caso della giornata di lunedì 24 aprile 2017 sono il topic identificato dall'algoritmo con il numero id *16* la cui *keyword* più rilevante è *Smack-Mellon* al livello di aggregazione 0.97, il topic con id *34* con *MFM* al livello di aggregazione 0.9 e il topic id *111* con *Miyonse* al livello di aggregazione 0.84. (fig. ☞ 86)
Questi tre topic mantengono la loro identità da"topic singolo" lungo la struttura gerarchica e soltanto quando tutti gli altri topic sono diventati cluster, si accorpano.



86



87

fig.86 I topic più forti del 24 aprile 2017 si accorpano ad altri cluster a livelli altissimi del dendrogramma (0.85, 0.9, 0.97).
fig.87 Schema esplicativo della clusterizzazione lungo il dendrogramma

Tuttavia la profondità di un topic può avere momenti di estrema debolezza e momenti di estrema robustezza dipendenti dalla definizione dei confini di quel determinato topic a quello specifico livello di dettaglio e analizzando il modo in cui questi topic si accorpano a diversi livelli di dettaglio, è possibile notare comportamenti diversi.

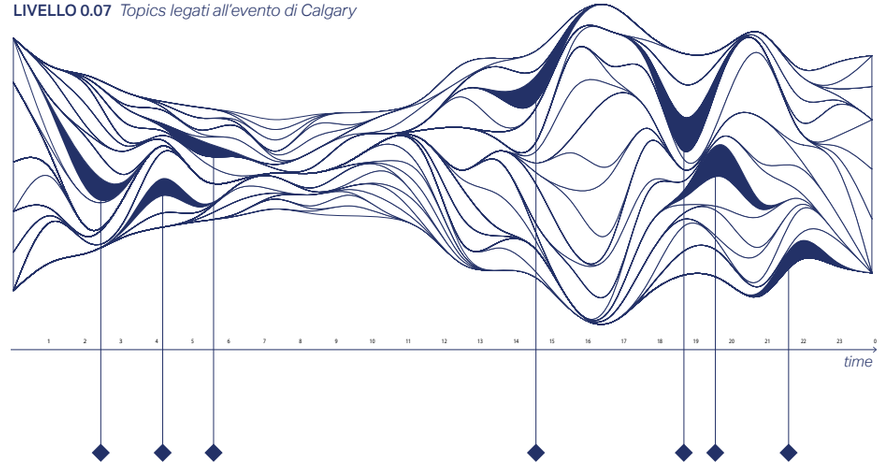
Preso come riferimento un topic presente a profondità 118 alle ore 21 che ha Calgary[77] come prima keyword caratterizzante si osserva come topic che trattano un argomento relativo a Calgary sono presenti anche in altri sei intervalli temporali già a partire dal livello di profondità 0.
Questo significa che su Twitter si è discusso di questo argomento prevalentemente in sette intervalli temporali: (2-3, 4-5, 5-6, 14-15, 18-19, 19-20, 21-22) ma, probabilmente in termini sempre diversi che rispecchiano la matrice di similarità. (fig. ☞ 89)
Già a profondità 0,08 riconosciamo un cluster di topic relativi all'argomento *Calgary* (fig. ☞ 90), indice di una apparente debolezza del topic. Tutti questi gruppi di *keywords* si riferiscono alla discussione emersa su Facebook e Twitter a seguito di un post di una donna, madre di un ragazzo morto di overdose a Calgary con l'obiettivo di sensibilizzare i giovani.

> *I just want everyone to know that my son Michael overdosed on fentanyl . My son was not an addict he made a mistake that cost him his life. I just want to make everyone aware of the epidemic that's goin on right now that's killing 5-7 people a day in every city in Canada. It's out of control and there is no way to protect our children from this other than to warn them of the dangers of drug use today.*
>
> — *S. Kent, 24 Aprile 2017*

77. Città dell'Alberta, Canada

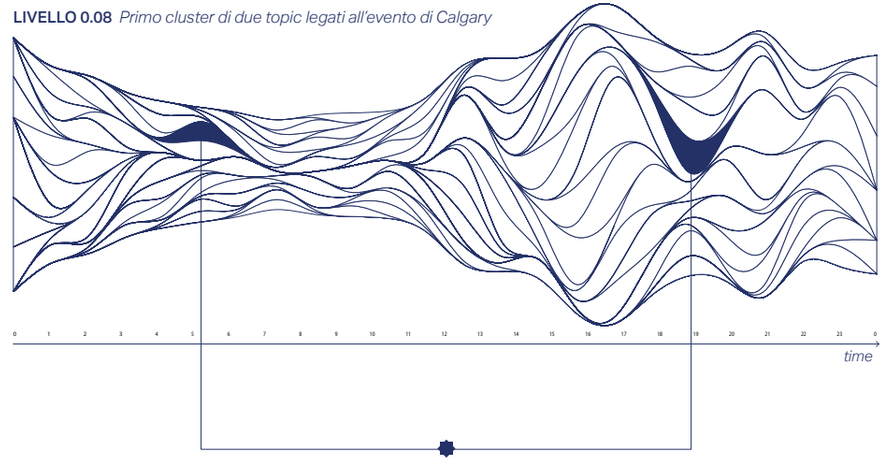
fig.88 Sherry Kent e suo figlio.

LIVELLO 0.07 *Topics legati all'evento di Calgary*



89

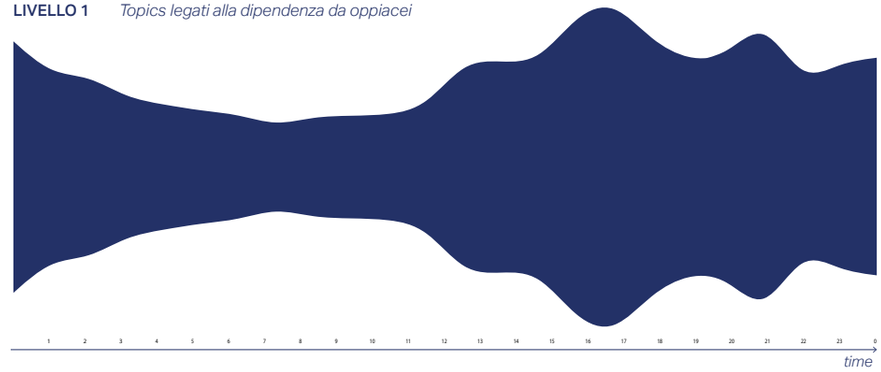
LIVELLO 0.08 *Primo cluster di due topic legati all'evento di Calgary*



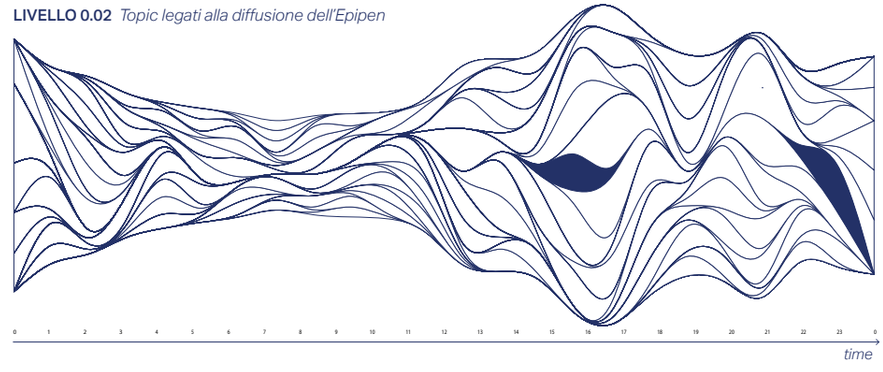
90

fig.89 Topic singoli legati all'evento di Calgary al livello 0.07 di profondità del dendrogramma.
fig.90 Al livello 0.08 del dendrogramma due dei topic singoli presenti al livello di profondità 0,07 diventano un cluster unico.

Quanto sono instabili i confini di un topic?
Presi i primi gruppi di *cluster* 120, 121 e 122 l'intento era quello di capire se fosse possibile definire l'*instabilità* di un topic partendo dal concetto di robustezza. (fig. ☞ 91)
Ogni topic oltre ad avere una sua evoluzione temporale ha anche la sua robustezza rispetto al grado di dettaglio in cui si osserva.
Un topic instabile è un topic che si ingloba all'interno di cluster più grandi e generici, un topic stabile è un topic che nonostante la profondità della struttura gerarchica non si accorpa a nessun altro per la maggior quantità di livelli possibile perché è talmente specifico che soltanto ai massimi livelli di generalizzazione è inglobato da cluster più grandi (come nel caso degli esempi precedenti di *Smack Mellon*, *MFM* e *Miyonse*). (fig. ☞ 89)
In *TopTom* la persistenza di un topic sul piano temporale è legata alla sua resistenza e stabilità a livello di struttura gerarchica, ovvero lungo il dendrogramma.
Per esempio nella giornata del 24 aprile 2017 il topic caratterizzato perlopiù dalla parola *addiction* (fig. ☞ 91) è molto resistente nel tempo, tanto da coprire tutte le 24 ore prese in analisi. Il topic caratterizzato dall'argomento della dipendenza dagli oppiacei è generico e, in qualche modo, padre di tutti i topic della giornata. Se invece si osserva lo *streamgraph* sfogliato a un livello più basso del dendrogramma si nota come l'argomento generico tenga a sparire e si caratterizzi in argomenti più specifici, come per esempio il topic 120 a livello 0,02 che raccoglie i tweets relativi all'utilizzo e alla vendita della *Epipen*, strumento utile per somministrare l'epinefrina.
In termini generici anche l'*Epipen* è un termine legato all'ampio campo della dipendenza da oppiacei, ma è solo grazie alla struttura gerarchica di clusterizzazione che è possibile analizzare queste differenze nel dettaglio e trovare trame multiple e differenti utili a modellare e monitorare l'argomento.(fig. ☞ 92 e 93)



91



92

fig.91 Cluster generico legato alla dipendenza da oppiacei.
fig.92 Il primo cluster della giornata, a livello 0.02 della struttura gerarchica del dendrogramma, mostra il tema della vendita dell'Epipen.

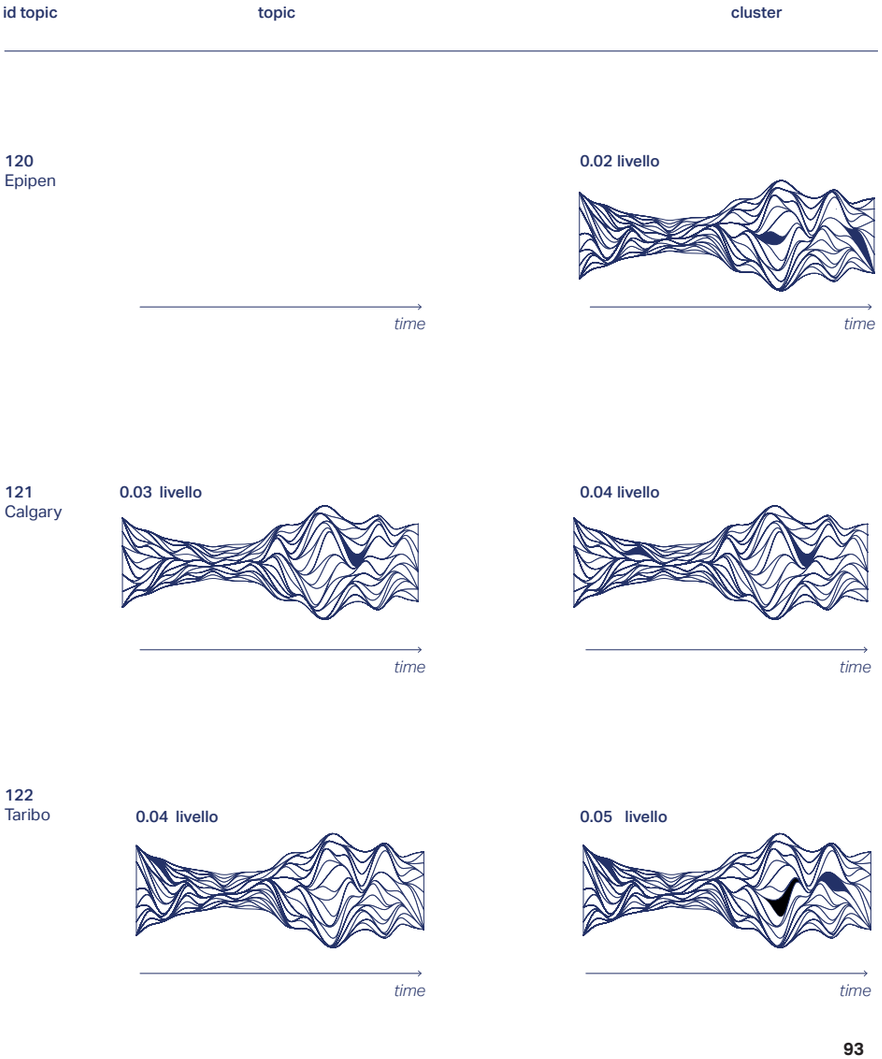


fig.93 Topic singoli legati all'evento di Calgary al livello 0.07 di profondità del dendrogramma.

Sebbene l'originalità di *TopTom* sia anche nella possibilità di vedere la gerarchia di clusterizzazione dei topic sfogliando lo *streamgraph* sull'asse z, dall'interfaccia non è possibile vedere le strutture gerarchiche a confronto in un'unica visualizzazione.
In *Hierarchical Topics* (W.Dou et al., 2013) fanno un tentativo, ma senza tenere conto della scarsa leggibilità poiché mantengono il modello visivo dello *streamgraph* come riferimento.
In *FluxFlow* (J. Zhao et al., 2014) per mostrare i livelli della struttura gerarchica è mantenuto il modello visivo di partenza (un *beeswarm*) e risulta difficile adattarsi al poco spazio dell'interfaccia.
Inoltre,a differenza di *TopTom*, sia *Hierarchical Topics* che *FluxFlow* sono piattaforme progettate per lo più come strumento di notazione, caratteristica che per il nostro progetto non era richiesta.

L'algoritmo di clusterizzazione *UPGMA* è l'ultimo tassello di tutto il processo algoritmico ed è quello che caratterizza fortemente questo tipo di *topic detection*.
Mettere in luce la struttura del dendrogramma su ogni topic può essere un risultato molto interessante ai fini della ricerca, cercando di progettare un modello visivo che possa essere applicato in futuro ad altri dati, magari frutto di altri algoritmi di clusterizzazione simili producendo risultati comparabili.

L'idea è quella di presentare una visualizzazione composta da un certo numero di elementi e l'obiettivo è quello di mettere in luce l'identità dei singoli topic sulla base della loro tendenza ad accorparsi
In altre parole, è la ricerca visiva dell'identità del topic secondo l'algoritmo di clusterizzazione *UPGMA*.

6.4.3 Identità forti e deboli in base alla struttura

Esistono quindi topic, come *Smack Mellon*, *MFM* e *Miyonse* che rimangono tali fino ai livelli di profondità più alti, altri che invece hanno bisogno di più passaggi prima di unirsi ad un topic generico nella parte alta del dendrogramma.
Quali di queste tipologia di topic ha un'identità forte?
Il topic 122 (fig. ☞ 93) che al massimo livello di dettaglio tratta un aspetto specifico del caso della madre canadese citato in precedenza, nel corso della sua generalizzazione mostra una condizione di stabilità nella zona centrale del dendrogramma, mentre si nota molto dinamismo nei pressi delle estremità.
Invece, il topic 120 è molto instabile nella zona più bassa del dendrogramma e si biforca in topic più specifici già a partire dal livello 0.54.
Per ottenere queste dati è stato creato un dataset originato dalle API in formato JSON.
La colonna id topic corrisponde al numero di riferimento del topic preso in considerazione, la colonna livello di cambio è il livello del dendrogramma a cui l topic cambia struttura e infine la colonna cluster indica l'id del topic nuovo, creatosi dall'unione di due topic, tra cui quello con l'id indicato dalla colonna topic.
Intenzione della visualizzazione non è trovare le relazioni tra i topic che si uniscono, ma definire l'identità del singolo topic nella sua struttura gerarchica.

Topic	Livello di cambio	Id topic	Id cluster
Epipen	118	120	120
Epipen	69	120	169
Epipen	65	120	173
Epipen	64	120	174
Epipen	29	120	209
Epipen	5	120	233
Epipen	2	120	236
Tari-taribo	117	121	121
Tari-taribo	67	121	171

6.5 Ristrutturare il topic con la visualizzazione

L' obiettivo è quindi visualizzare l'identità di ogni topic sulla base del concetto di stabilità e instabilità della sua struttura al fine di poter confrontare i comportamenti a diversi livelli di stabilità nell'arco di uno specifico arco temporale ed evidenziare eventuali similarità anche soltanto osservando la visualizzazione.
La rappresentazione dell'instabilità nella struttura del topic è già chiara nel dendrogramma ma non si evince facilmente la similarità di contenuto tra i topic.
Se si considerano su una linea che va dal livello 0 al livello 1 i cambiamenti di struttura di ogni singolo topic si possono ottenere diverse conformazioni che mostrano la frequenza (in termini di formazione di cluster) con cui quel determinato topic cambia aspetto, accorpandosi a topic simili.
L'identità del topic può essere per cui definita a partire dal suo essere stabile o instabile, come se fosse una struttura.
Una struttura può essere debole e instabile alle sue fondamenta (livello 0 del dendrogramma) o durante molti livelli di profondità; una struttura può invece essere definita stabile se mantiene una struttura lineare lungo un ampia zona (verticale) del dendrogramma.
Prima di definire il modello visivo è tuttavia necessario comprendere quale sia il miglior modo per rappresentare i topic in una visione olistica, in modo che siano facilmente comparabili e non occupino molto spazio.

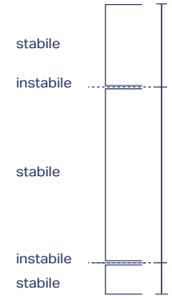


fig.94 Struttura gerarchica. Con riferimenti alle zone di instabilità.

Nel capitolo dedicato allo *small multiples*, E. Tufte sostiene che lo *small multiples*[78] sia il modo migliore per enfatizzare i cambiamenti al livello di dato.

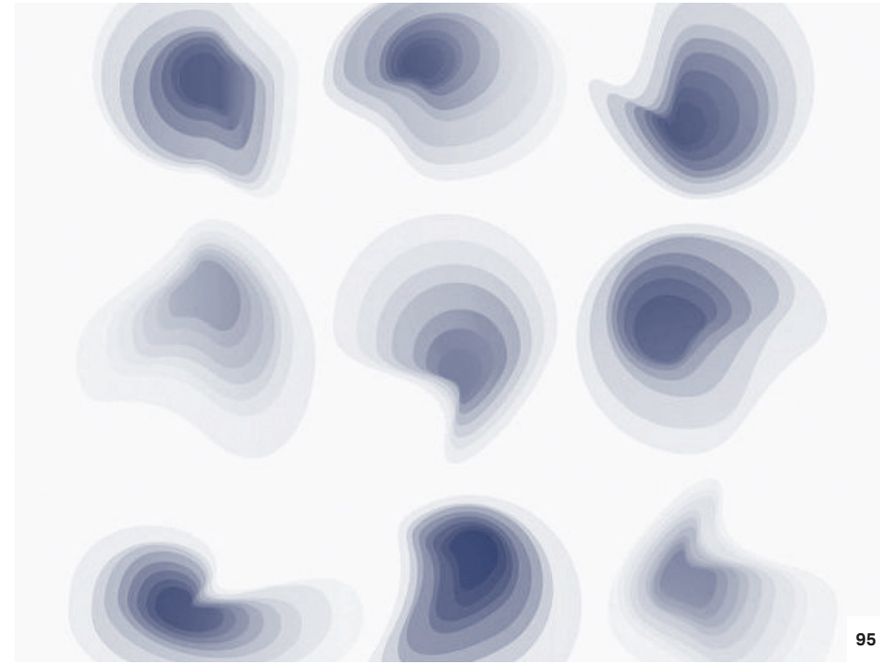
> *Constancy of design puts the emphasis on changes in data, for a wide range of problems in data presentation, small multiples are the best design solution.*
>
> — *E. Tufte, 1991*

Questa sperimentazione, oltre a voler enfatizzare il cambiamento del dato, vuole enfatizzare anche il cambiamento della struttura del dato.
Inoltre, poiché l'obiettivo è mettere in luce le multiple identità dei topic attraverso l'instabilità delle strutture è interessante considerare anche che

> *Small multiple reveals, all at once, a scope of alternatives, a range of options.*
>
> — *E. Tufte, 1991*

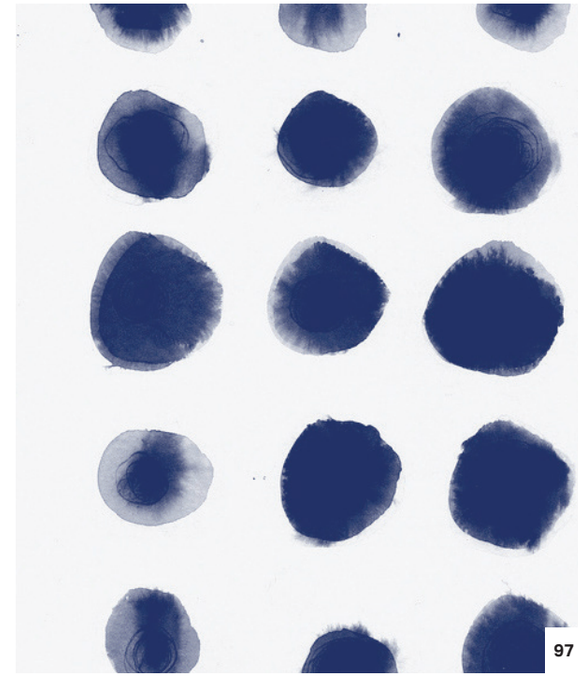
La scelta di rappresentare il topic come forma circolare deriva dalla necessità di dare aspetto dimensionale alla linea/raggio della pagina precedente sia in *TopTom* con il modello visivo *metaball*, che in *Termite* (J. Chuang et al., 2012) il topic è rappresentato come un elemento circolare. Inoltre, la necessità di mostrare l'identità del topic simula un processo di schedatura dei topic per identificarne similarità e differenze, come nel caso delle impronte digitali di più individui a confronto.
(fig. ☞ 98 e 99)

78. Serie di grafici simili che usano stessa scala e assi.



95

96



97

fig.95 Gradient Tomography, 2010
fig.96 Impronte digitali
fig.97 Fotografia, E.Nanni, 2013

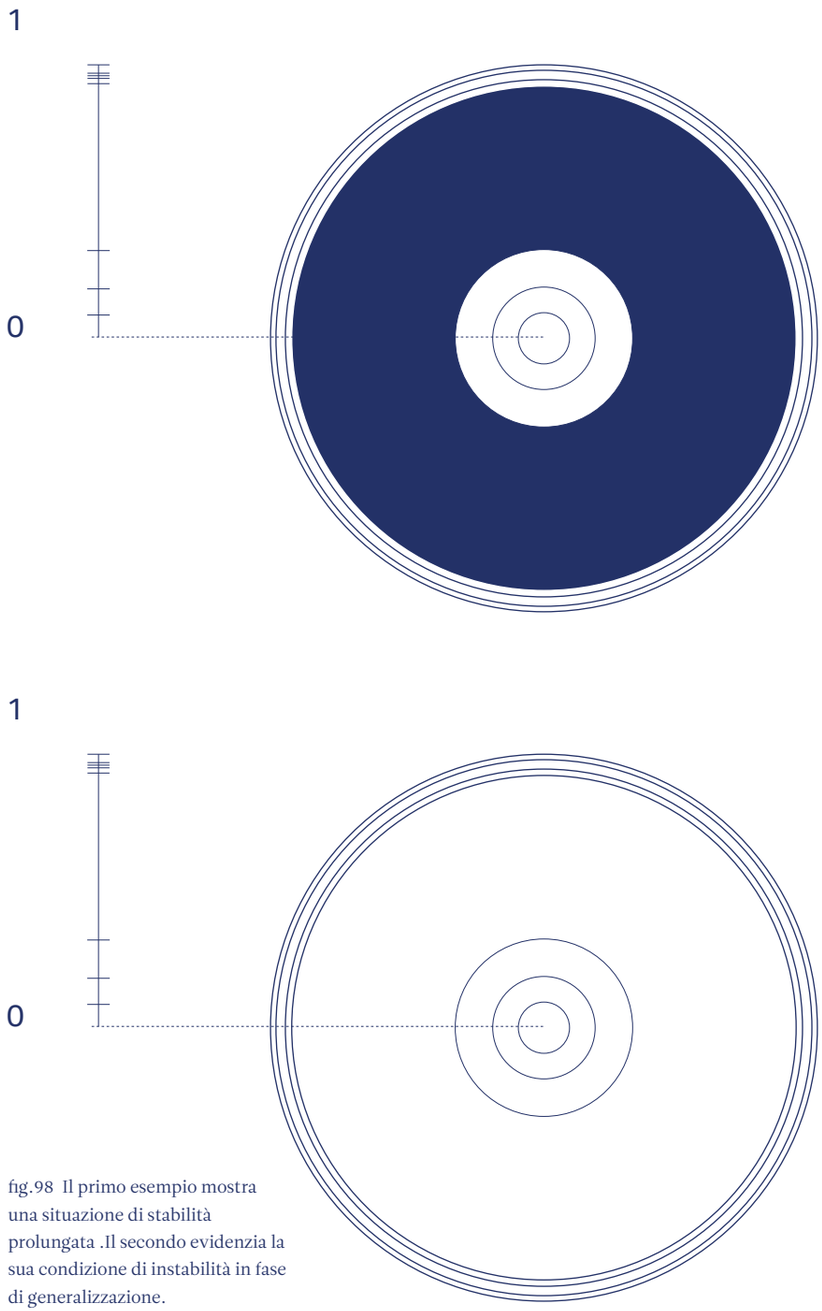
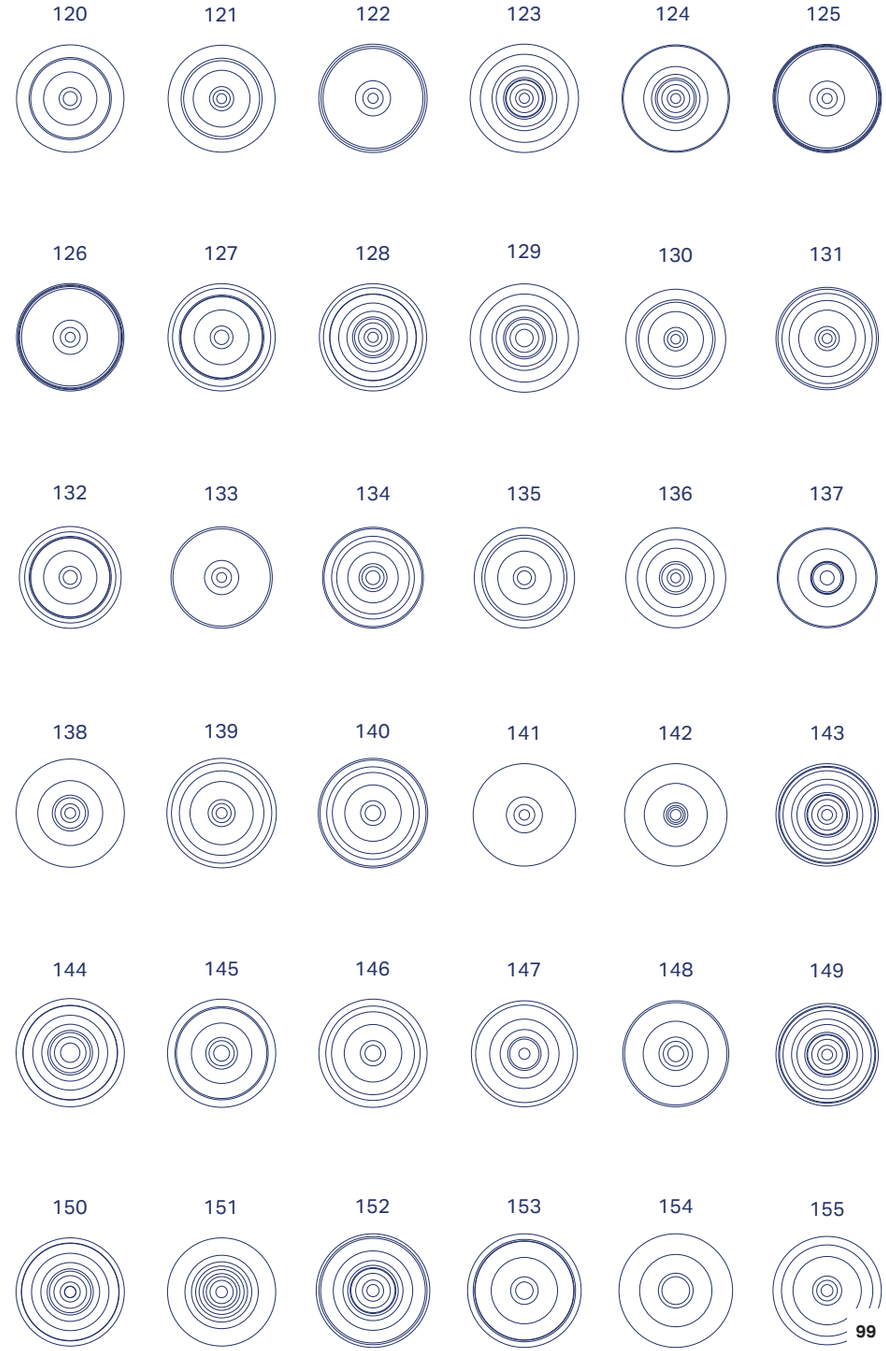
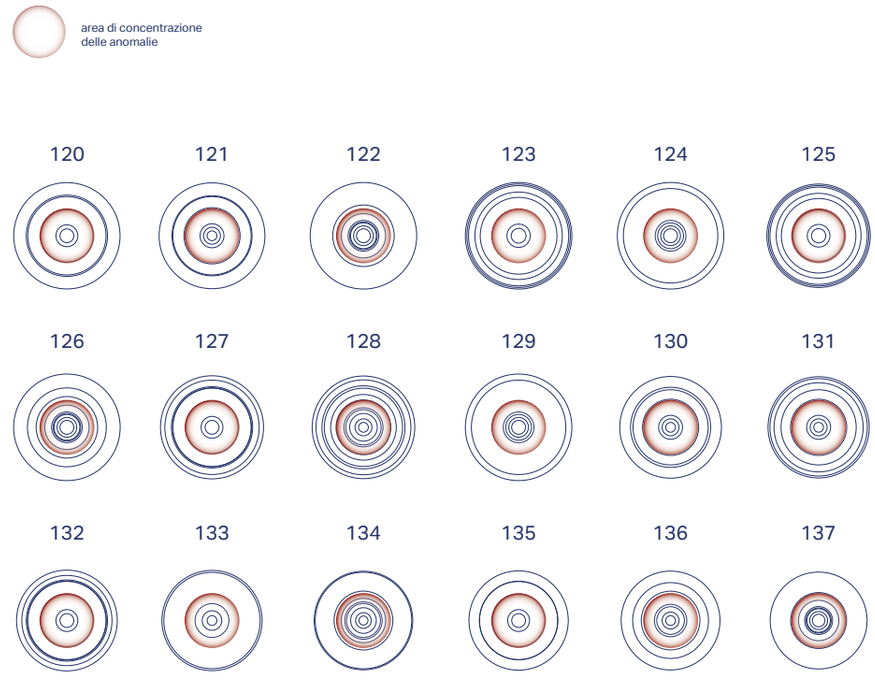


fig.98 Il primo esempio mostra una situazione di stabilità prolungata .Il secondo evidenzia la sua condizione di instabilità in fase di generalizzazione.
fig.99 Un esempio di *small multiples* dei primi 36 topic.



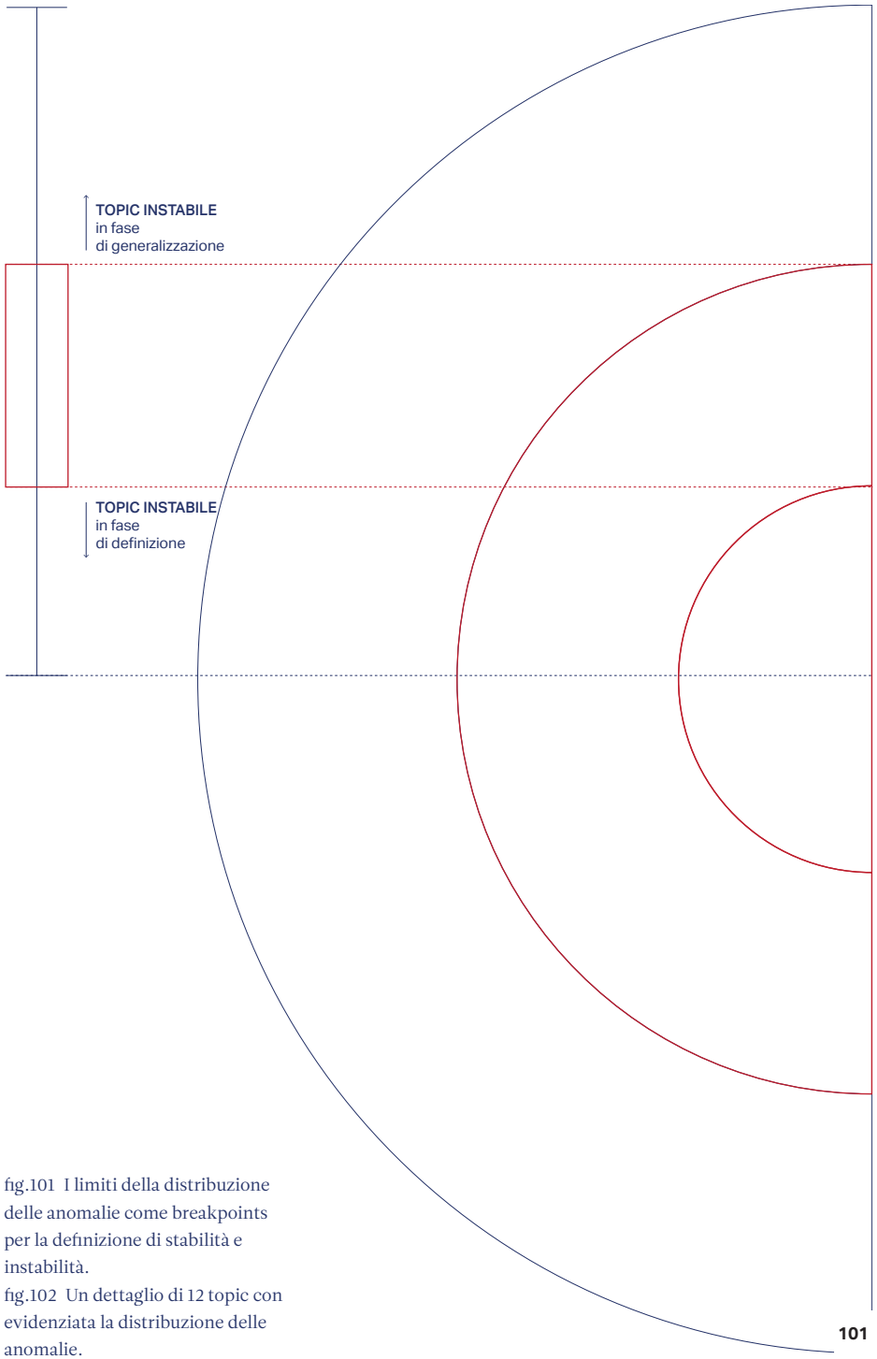
6.6 Mantenere la complessità

Esistono topic più stabili di altri e da cosa può dipendere la loro stabilità? E, inoltre, è possibile trovare un modello visivo in grado di dare informazioni sull'identità del topic e facilitare la comprensione del funzionamento dell'algoritmo di clusterizzazione senza partire dalla struttura dei contenuti?
La stabilità di un topic è una caratteristica interessante che può essere rappresentata dalla frequenza di mutazioni di un topic lungo la sua struttura gerarchica.
In questo modo, *la struttura del dato guida il modello visivo senza ledere l'identità e la complessità del topic* perché non si basa solo sulla rappresentazione del contenuto ma sulla rappresentazione del comportamento permettendo di vedere in una sola vista una caratteristica del dataset che con *TopTom* non era possibile osservare nel complesso.
Da questa analisi è stato possibile identificare tre diverse tipologie di comportamento del topic.
Già a vista d'occhio si può notare come alcuni topic tendano a definirsi in fase di aggregazione iniziale, altri cambino in maniera costante ed altri ancora siano instabili in fase di definizione generale. Per determinare tali breakpoints può essere interessante utilizzare lo spettro delle anomalie che si concentrano, per la giornata di tweets del 24 Aprile, tra il livello 0.3 e 0.6 della gerarchia, quindi in posizione centrale.
(fig. ☞ 100, 101 e 102)



100

fig.100 In alto i primi 18 topic con i breakpoints evidenziali.



101

fig.101 I limiti della distribuzione delle anomalie come breakpoints per la definizione di stabilità e instabilità.
fig.102 Un dettaglio di 12 topic con evidenziata la distribuzione delle anomalie.

120

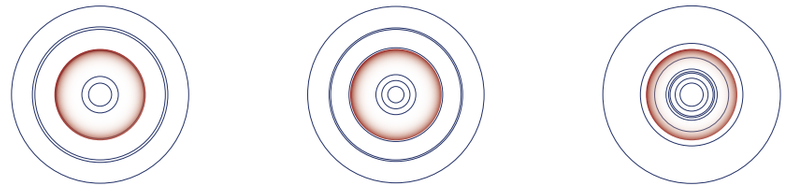
To keep EpiPen sales up, Mylan threatened states, sued making bogus claims

121

Senator Isiaka Adeleke, Died Of An Overdose Of Painkillers

122

Calgary mother hopes photo of dying son will deter others from doing fentanyl



123

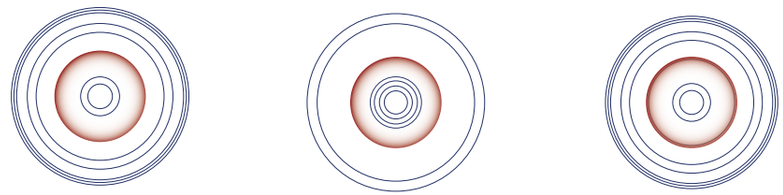
Cherokee_Nation files lawsuit targeting CVS and other pharmacies

124

This is really outstanding W5's '48 Hours' investigation: Lessons from Vancouver

125

Stop giving codeine and Tramadol to Children_–_FDA warns - the food and drug administration says ...



126

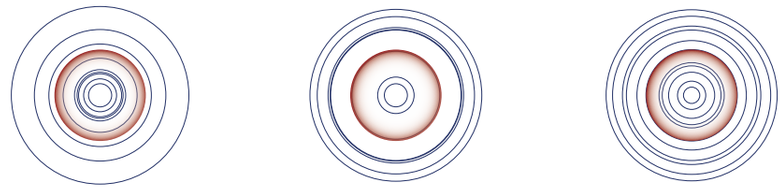
Calgary mother hopes photo of dying son will deter others from doing fentanyl -

127

Mylan took one state to court to push EpiPen sales, documents show

128

Trump 'hate-watches' news that is critical



129

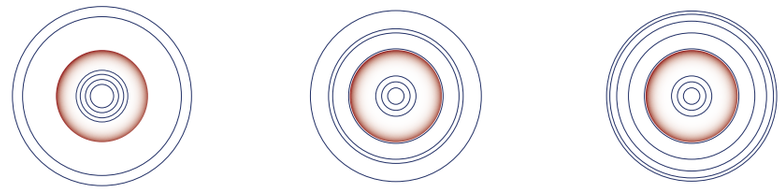
Calgary mother hopes photo of dying son will deter others from doing fentanyl -

130

Senator Isiaka_Adeleke died of overdose of painkillers (Details of Autopsy)

131

Senator Adeleke died of drug overdose



102

Per cui, sulla base dei breakpoints originati dalle anomalie possiamo definire tre categorie di struttura del topic. (fig. ☞ 102)

☞ Topic instabile in fase di definizione (instabilità osservata prima del livello 0.3) Accade con documenti molto specifici legati ad un evento di cui si è parlato a lungo nel tempo.

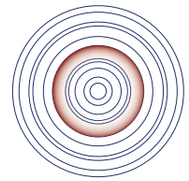
☞ Topic con costante instabilità. (instabilità osservata lungo molti livelli della struttura gerarchica) Accade con topic mutevoli, che sebbene trattino di argomenti specifici inizialmente poi sono subito assimilabili ad altri. Spesso, come nel caso del topic 128 trattano temi inerenti a persone implicate in più discussioni.

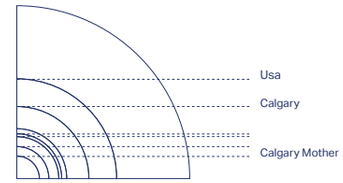
☞ Topic instabile in fase di generalizzazione (instabilità osservata o dopo il livello 0.6). Accade con documenti molto specifici legati ad un evento di cui si è parlato solo in un momento della giornata.

Quest'ultimo capitolo ha dato la possibilità di esplorare le potenzialità del dato ed aperto le porte a nuove sperimentazioni.
Si potrebbe osservare se è possibile ottenere un risultato interessante considerando come parametro chiave il contenuto, le relazioni o il tempo ed, eventualmente confrontare i risultati.

TOPIC INSTABILE
lungo tutta la struttura

128

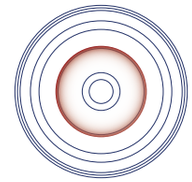


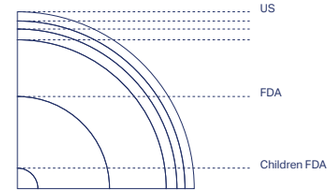


Trump 'hate-watches'
news that is critical

TOPIC INSTABILE
in fase di generalizzazione

125

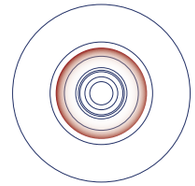


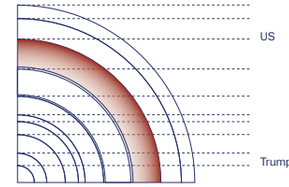


Stop giving codeine
and Tramadol to
Children_–_FDA

TOPIC INSTABILE
in fase di generalizzazione

122





Calgary mother
hopes photo of dying
son will deter others
from doing fentanyl

Su di un cerchio ogni punto d’inizio può anche essere un punto di fine.

— Eraclito, 400 a.C

7. Il contributo di *Tomotopigrafie*

Questa tesi nasce con l'obiettivo di applicare conoscenze e competenze di ricerca del designer dell'informazione alla visualizzazione dei processi e dei risultati di algoritmi di *topic modeling* dinamico e gerarchico, argomento che fino a qualche anno fa era ristretto a soli statistici e data scientists.
Durante la progettazione del tool interattivo e nel corso della sperimentazione finale sono emersi alcuni punti chiave.

Sul ruolo del designer

☞ Trattandosi di un neonato ambito di ricerca dal punto di vista del design, è necessario analizzare una bibliografia ibrida, considerando sia articoli di matrice statistico/matematica riguardanti innovazioni algoritmiche legate al *topic modeling* che progetti di visualizzazione dati e design dell'informazione.

☞ Dalla ricerca è emerso come ogni *topic detection*, caratterizzata da una serie di lunghezza variabile di algoritmi specifici, richieda modelli visivi ad hoc, per questo motivo il dialogo tra designer e *data scientist* è di vitale importanza ai fini della comprensione e della progettazione della struttura annidata dei risultati.

Sulla rappresentazione del topic

☞ Il topic è un'entità astratta e complessa, che esiste solo quando definita da altri parametri (*keywords*, tempo, volume, profondità...), per questo motivo è importante trattare la sua rappresentazione con cautela, e non semplificarne in maniera sprovveduta il contenuto.

☞ Poiché ogni serie di algoritmi di *topic modeling* necessita di modelli visivi *ad hoc*, allo stesso modo anche la rappresentazione del topic è strettamente correlata al tipo di struttura dei metadati che lo definiscono nei risultati delle analisi. Nel capitolo 6 viene dato particolare risalto al parametro della gerarchia, punto di forza e innovazione dei risultati dell'algoritmo su cui è stata effettuata la ricerca.

☞ Nella rappresentazione complessa dell'identità del topic è la struttura gerarchica del dato a guidare il modello visivo perché consente di visualizzare il comportamento del topic.

Sulla scelta dei modelli visivi

☞ La letteratura mostra come ogni innovazione algoritmica abbia richiesto anche innovazioni di modello visivo. I più recenti algoritmi della famiglia del *topic modeling* sfruttano la tridimensionalità dello spazio per definire il dato, allo stesso modo, immaginare modelli visivi noti come lo *streamgraph* in uno spazio

tridimensionale, può aiutare a simulare nella maniera più affidabile possibile la struttura del dato.

☞ La ricerca e l'utilizzo di una metafora visiva in grado di adattarsi a viste multiple coordinate facilita sia la rappresentazione che l'interpretazione dei risultati delle analisi. Nel caso di *TopTom* la metafora della TAC, appartenente all'ambito medico-scientifico, è dominante, ed ogni elemento aggiuntivo è legato alla metafora stessa come nel caso del liquido di contrasto usato per rendere le anomalie evidenti.

☞ L'efficienza di funzionamento di un modello visivo è correlata non solo al contenuto, ma alla struttura stessa del dataset, così come l'efficienza di funzionamento del tool è garantita dall'efficacia simultanea di modello visivo e interfaccia.

☞ La ricerca presentata è da considerarsi solo un punto di partenza per ulteriori sperimentazioni nell'ambito della visualizzazione dei risultati di algoritmi di *topic modeling* dinamico-gerarchico, sia per quanto riguarda i limiti legati alla rappresentazione coordinata e multi vista (cap. ☞ 5) che per quelli ristretti alla rappresentazione del topic (cap. ☞ 6).

☞ Lo studio e la visualizzazione del topic proposti nel capitolo 6 potrebbero sia essere inseriti all'interno di una rappresentazione coordinata e multi-vista, in ottica di migliorare la fruizione del dato e fornire informazioni sul comportamento del topic in termini di struttura gerarchica, ma potrebbero anche diventare un ipotetico rapido strumento di analisi preliminare della tipologia di tematica del topic che non dipenda dal suo contenuto ma solo dalla sua struttura.

La speranza è che questo lavoro abbia contribuito ad aggiungere un tassello mancante alla ricerca di design e che possa servire ad altri studenti o ricercatori come punto di inizio per lavorare in contesti simili a quello analizzato, fornendo una visione forse un po' più chiara dell'argomento.

8. Bibliografia e sitografia

A

Abbott, E, *Flatlandia, Racconto fantastico a più dimensioni*, Gli Adelphi, Milano (1966)

Alexander, E. & Gleicher, M. *Task-Driven Comparison of* Topic *Models*. IEEE Trans. Vis. Comput. Graph. 22, 320–329 (2016).

Alighieri D., La Divina Commedia - Inferno

Aristotele, *Organon*, Topici, 408-439, Giulio Einaudi Editore, 1955

B

Baldonado, M. Q. W., Woodfruss, A. & Kuchinsky, A. *Guidelines for using multiple views in information visualization.* AVI '00 Proc. Work. Conf. Adv. Vis. Interfaces 110–119 (2000). doi:10.1145/345513.345271

Baule, G., Caratti, E., Design è traduzione. Il paradigma traduttivo per la cultura del progetto. «Design e traduzione»: un manifesto, Franco Angeli Editore, Milano, 2016

Baur, D., Lee, B. & Carpendale, S. *TouchWave: Kinetic Multi-touch Manipulation for Hierarchical Stacked Graphs*. Proc. 2012 ACM Int. Conf. Interact. tabletops surfaces 255–264 (2012). doi:10.1145/2396636.2396675

Bernstein, M. S. et al. *Eddi: Interactive* Topic*-based Browsing of Social Status Streams*. Proc. 23nd Annu. ACM Symp. User interface Softw. Technol. - UIST '10 303 (2010). doi:10.1145/1866029.1866077

Bertin, J. & Les, C. *Sémiologie graphique*. Image (Rochester, N.Y.) 7–10 (2009). doi:10.1037/023518

Blei, D. M. et al. *Latent Dirichlet Allocation*. J. Mach. Learn. Res. 3, 993–1022 (2003).

Blei, D. M., Griffiths, T. L. & Jordan, M. I. *The nested Chinese restaurant process and Bayesian nonparametric inference of* topic *hierarchies*. 57, 1–30 (2007).

Blei, D. M. & Lafferty, J. D. *Dynamic* topic *models*. Proc. 23rd Int. Conf. Mach. Learn. - ICML '06 113–120 (2006). doi:10.1145/1143844.1143859

Blei, D., Carin, L. & Dunson, D. *Probabilistic* topic *models*. IEEE Signal Process. Mag. 27, 55–65 (2010).

Byron, L. & Wattenberg, M. *Stacked graphs - Geometry & aesthetics*. IEEE Trans. Vis. Comput. Graph. 14, 1245–1252 (2008).

C

Cao, N. et al. Whisper: *Tracing the Spatiotemporal Process of Information Diffusion in Real Time*. IEEE Trans. Vis. Comput. Graph. 18, 2649–2658 (2012).

Chaney, A. & Blei, D. *Visualizing* Topic *Models*. Icwsm 419–422 (2012).

Choo, J., Lee, C., Reddy, C. K. & Park, H. *UTOPIAN : User-Driven Topic modeling Based on Interactive Nonnegative Matrix Factorization*. 19, 1992–2001 (2013).

Chuang, J., Jin, A., Mcfarland, D. A., Wilkerson, J. D. & Manning, C. D. Document Exploration with *Topic modeling* : *Designing Interactive Visualizations to Support Effective Analysis Workflows*. 1–4

Chuang, J., Manning, C. D. & Heer, J. *Termite : Visualization Techniques for Assessing Textual* Topic *Models*. Proc. Int. Work. Conf. Adv. Vis. Interfaces - AVI '12 74 (2012). doi:10.1145/2254556.2254572

Chuang, J., Ramage, D., Manning, C. D. & Heer, J. *Interpretation and Trust : Designing Model-Driven Visualizations for Text Analysis*. (2005).

Collins, C., Viégas, F. B. & Wattenberg, M. *Parallel tag clouds to explore and analyze faceted text corpora*. VAST 09 - IEEE Symp. Vis. Anal. Sci. Technol. Proc. 91–98 (2009). doi:10.1109/VAST.2009.5333443

Cui, W., Liu, S., Wu, Z. & Wei, H. *How Hierarchical Topics Evolve in Large Text Corpora*. 20, 2281–2290 (2014).

Cui, W. et al. *TextFlow: Towards Better Understanding of Evolving* Topic *in Text*. IEEE Trans. Vis. Comput. Graph. 17, 2412–2421 (2011).

D

Deerwester, S., Dumais, S. T. & Harshman, R. *Indexing by latent semantic analysis*. J. Am. Soc. Inf. Sci. 41, 391–407 (1990).

Donath, J., Karahalios, K. & Viégas, F. *Visualizing Conversation*. J. Comput. Commun. 4, 0–0 (2006).

Dörk, M., Gruen, D., Williamson, C. & Carpendale, S. *A Visual Backchannel for Large-Scale Events*. IEEE transactions on visualization and computer graphics 16, (2011).

Dou, W., Wang, X., Skau, D., Ribarsky, W. & Zhou, M. X. *LeadLine: Interactive Visual Analysis of Text Data through Event Identifcation and Exploration*. Proc. 2012 IEEE Conf. Vis. Anal. Sci. Technol. 93–102 (2012). doi:10.1109/VAST.2012.6400485

Dou, W., Yu, L., Wang, X., Ma, Z. & Ribarsky, W. *Hierarchical*Topic*: Visually*

exploring large text collections using topic *hierarchies*. IEEE Trans. Vis. Comput. Graph. 19, 2002–2011 (2013).

E

Eco, U., *Come si fa una tesi di laurea*, Milano, Bompiani, 1977.

F

Finn, E., *What Algorithms Want - Imagination in the Age of Computing*, MIT Press, (2017)

G

Glueck, M. et al. *PhenoLines: Phenotype Comparison Visualizations for Disease Subtyping via* Topic *Models*. IEEE Trans. Vis. Comput. Graph. 24, 371–381 (2018).

H

Havre, S., Hetzler, E. & Nowell, L. *ThemeRiver: In Search of Trends, Patterns, and Relationships*. InfoVis 99 4 (1999). doi:10.1109/INFVIS.2000.885098

Hofmann, T. *Probabilistic Latent Semantic Analysis*. Proc. Fifteenth Conf. Uncertain. Artif. Intell. 289–296 (1999). doi:10.1.1.33.1187

K

Kim, M., Kang, K., Park, D., Choo, J. & Elmqvist, N. Topic*Lens: Efficient Multi-Level Visual* Topic *Exploration of Large-Scale Document Collections*. IEEE Trans. Vis. Comput. Graph. 23, 151–160 (2017).

1. Kucher, K. & Kerren, A. *Text visualization techniques: Taxonomy, visual survey, and community insights*. IEEE Pacific Vis. Symp. 2015–July, 117–121 (2015).

L

Lancichinetti, A. & Fortunato, S. *Consensus clustering in complex networks*. Sci. Rep. 2, 336 (2012).

Leskovec, J., Backstrom, L. & Kleinberg, J. *Meme-tracking and the dynamics of the news cycle*. Proc. 15th ACM SIGKDD Int. Conf. Knowl. Discov. data Min. - KDD '09 497 (2009). doi:10.1145/1557019.1557077

Liu, S. et al. Topic*Panorama : a Full Picture of Relevant* Topic. 183–192 (2014).

Liu, S. et al. *StoryFlow : Tracking the Evolution of Stories*. 19, 2436–2445 (2013).

M

Maceachren, A. M., Roth, R. E., O'Brien, J., Swingley, D. & Gahegan, M. *Visual Semiotics and Uncertainty Visualisation: An Empirical Study*. IEEE Trans. Vis. Comput. Graph. 18, 1–10 (2012).

Malik, S. et al. Topic*Flow: Visualizing* Topic *Alignment of Twitter Data over Time*. Proc. 2013 IEEE/ACM Int. Conf. Adv. Soc. Networks Anal. Min. - ASONAM '13 720–726 (2013). doi:10.1145/2492517.2492639

Mauri, M., Pini, A., Ciminieri, D. & Ciuccarelli, P. *Weaving data, slicing views: A design approach to creating visual access for digital archival collections*. ACM Int. Conf. Proceeding Ser. CHItaly 1–8 (2013). doi:10.1145/2499149.2499159

P

Panisson, A., Gauvin, L., Quaggiotto, M. & Cattuto, C. *Mining concurrent* topic*al activity in microblog streams*. CEUR Workshop Proc. 1141, 3–10 (2014).

R

Rönnqvist, S., Wang, X. & Sarlin, P. Interactive Visual Exploration of Topic Models using Graphs. (2014).

Rose, S., Butner, S., Cowley, W., Gregory, M. & Walker, J. Describing story evolution from dynamic information streams. VAST 09 - IEEE Symp. Vis. Anal. Sci. Technol. Proc. 99–106 (2009). doi:10.1109/VAST.2009.5333437

S

Salton, G., Wong, A. & Yang, C. S. *A vector space model for automatic indexing*. Commun. ACM 18, 613–620 (1975).

Scagnetti, G., Ricci, D., Baule, G. & Ciuccarelli, P. *Reshaping communication design tools. Complex systems structural features for design tools*. Proc. IASDR 07 (2007).

Shi, C. L. et al. *RankExplorer: Visualization of Ranking Changes in Large Time Series Data*. IEEE Trans. Vis. Comput. Graph. 18, 2669–2678 (2012).

Shi, L. et al. *Understanding text corpora with multiple facets*. VAST 10 - IEEE Conf. Vis. Anal. Sci. Technol. 2010, Proc. 99–105 (2010). doi:10.1109/VAST.2010.5652931

Sokal, R. R. & Michener, C. D. A Statistical Method for Evaluating Systematic Relationships. The University of Kansas Science Bulletin 38, (1958).

Sun, G. et al. *EvoRiver: Visual analysis of* topic *coopetition on social media. IEEE Trans*. Vis. Comput. Graph. 20, 1753–1762 (2014).

T

Tufte, E., *Envisioning Information*. Bull. Med. Libr. Assoc. 79, 346–348 (1991).

V

Venturini, T., Bounegru, L., Jacomy, M. & Gray, J. *How to Tell Stories with Networks: Exploring the Narrative Affordances of Graphs with the Iliad*. datafied Soc. Stud. Cult. through data 1–13 (2015). doi:10.5117/9789462981362

Viégas, F. B., Golder, S. & Donath, J. *Visualizing email content*. Proc. SIGCHI Conf. Hum. Factors Comput. Syst. - CHI '06 979 (2006). doi:10.1145/1124772.1124919

Viégas, F. B. & Wattenberg, M. *Tag clouds and the case for vernacular visualization*. Interactions 15, 49 (2008).

Viégas, F. & Wattenberg, M. *Google+ ripples: A native visualization of information flow*. Proc. WWW 2013 1389–1398 (2013).

W

Wang, X. et al. Topic*Panorama: A Full Picture of Relevant* Topic. IEEE Trans. Vis. Comput. Graph. 22, 2508–2521 (2016).

Wang, X. & McCallum, A. Topic *over time: A non-Markov continuous-time model of* topic*al trends*. Proc. 12th ACM SIGKDD Int. Conf. Knowl. Discov. Data Min. 424–433 (2006). doi:10.1145/1150402.1150450

Wei, F. et al. *TIARA: a visual exploratory text analytic system*. Kdd '10 153–162 (2010). doi:10.1145/1835804.1835827

53. Wu, Y., Liu, S., Yan, K., Liu, M. & Wu, F. *OpinionFlow: Visual analysis of opinion diffusion on social media*. IEEE Trans. Vis. Comput. Graph. 20, 1763–1772 (2014).

X

Xu, P., Wu, Y., Wei, E. & Peng, T. Visual Analysis of Topic Competition on Social Media. 19, 2012–2021 (2013).

Y

Yang, Y., Wang, J., Huang, W. & Zhang, G. Topic*Pie: An interactive visualization for LDA-based* topic *analysis*. Proc. - 2016 IEEE 2nd Int. Conf. Multimed. Big Data, BigMM 2016 25–28 (2016). doi:10.1109/BigMM.2016.25

Zhao, J. et al. *#FluxFlow: Visual analysis of anomalous information spreading on social media*. IEEE Trans. Vis. Comput. Graph. 20, 1773–1782 (2014).

Alexa Top Sites
https://www.alexa.com/topsites

D3 Library
https://d3js.org/

David Blei lectures on Topic *Models.*
https://www.youtube.com/watch?v=DDq3OVp9dNA&t=1778s.
https://www.youtube.com/watch?v=FkckgwMHP2s

Cascade, New York Times Labs, 2011
http://nytlabs.com/projects/cascade.html

Classic Jazz Song
http://kyrandale.com/viz/static/expts/d3-jazz/index.html

Data Visualization, Wikipedia page
https://en.wikipedia.org/wiki/Topic_model

Emoto Topic *Explorer*, Data Interfaces, 2012
http://www.datainterfaces.org/projects/emoto/

Emoto Instsllation, Studio Nand, 2012
http://www.nand.io/projects/clients/emoto-installation/

EvoRiver, S. Guodao, 2015
https://www.youtube.com/watch?v=MP0_L2Zn3eg&t=337s&list=PLZ84VWVYf-GFAj8fU_lS1Kkdj6ufy-v-vs&index=1

FluxFlow, J. Zhao, 2014
https://www.youtube.com/watch?v=OZMubJ0v32Q&t=8s

Galaxy of Covers, Interactive Things
https://lab.interactivethings.com/galaxy-of-covers/

How riot rumours spread on Twitter, R. Procter et al., 2011
https://www.theguardian.com/uk/interactive/2011/dec/07/london-riots-Twitter

Kepler, Alexis Lloyd, 2013
http://www.alexislloyd.com/projects/kepler.html

Kindred Britain, N. Jenkins, E. Meeks and S. Murray, 2013
http://kindred.stanford.edu/#

LDA Topic *Models*
https://www.youtube.com/watch?v=3mHy4OSyRf0&t=601s

Matrix factorization
https://www.youtube.com/watch?v=o8PiWO8C3zs&t=281s

Opinion Flow, Microsoft Research center, 2015
https://www.youtube.com/watch?v=aeEmbv6XRFc&list=PLZ84VWVYfGFAj8fU_lS1Kk-dj6ufy-v-vs&index=3&t=56s

Parallel Tag Clouds, C. Collins, F. Viegas, 2009
https://www.youtube.com/watch?time_continue=1&v=rL3Ga6xBgLw

Quantify Kissinger
quantifyingkissinger.com/stack-stack.html

*Rev*Isit, M. Stefaner, 201
http://truth-and-beauty.net/projects/revIsit/

Singular Value Decomposition (SVD)
https://www.youtube.com/watch?v=mBcLRGuAFUk

StoryFlow, L. Shixia, 2013
https://www.youtube.com/watch?v=yoq82mC30Iw&list=PLZ84VWVYfGFAj8fU_lS1Kkdj6ufy-v-vs&index=2

Tensor decomposition with Python: Learning structures from multidimensional data
https://www.youtube.com/watch?v=YuB2exVzd1s&t=552s

Text Visualization Browser
http://textvis.lnu.se/

The Free Dictionary
http://www.thefreedictionary.com/

The Rithm of Food, M. Stefaner, 2016
http://truth-and-beauty.net/projects/the-rhythm-of-food

TOP 2000 love song the 70's & 80's
http://www.datasketch.es/december/code/nadieh/

Topic *Model,* Wikipedia page
https://en.wikipedia.org/wiki/Topic_model

Topic *Flow*, S. Malik, A. Smith, T. Hawes, P. Papadatos, J Li; University of Maryland, Computer Science, 2013
http://www.cs.umd.edu/~maliks/topicflow/TopicFlow.html#

Termite, Stanford, 2012
http://vis.stanford.edu/topic-diagnostics/model/silverStandards/

Trump Connections, K. Albrecht, 2016
http://trump.kimalbrecht.com/#11

Utopian, J. Choo, 2013
https://www.youtube.com/watch?v=du6_s6hcaRA&index=4&list=PLZ84VWVYf-GFAj8fU_lS1Kkdj6ufy-v-vs

Visualizing Algorithms
https://bost.ocks.org/mike/algorithms/

Wahl 2017, M. Stefaner, 2017
http://truth-and-beauty.net/projects/wahl-2q17

9. Ringraziamenti

Il ringraziamento più grande va a mamma e papà, per avermi sostenuto in questi anni, sia moralmente che economicamente.
A Paolo, relatore di questa tesi, che ha creduto in me sin dall'inizio dandomi l'opportunità di lavorare a questo stimolante progetto.
A Michele, guida fondamentale durante questi mesi, per i suoi preziosi consigli, per la sua disponibilità e per avermi insegnato che il codice, talvolta, va preso a martellate.
Agli amici di Density Design che hanno reso questo percorso una delle esperienze formative più belle della mia vita: ad Àngeles, per il suo continuo sostegno nonostante gli impegni del dottorato, a Gabriele, per avermi aiutato ad aprire gli occhi in più di un'occasione, a Mitch, per l'attenzione con cui mi ha sempre ascoltata, e a Tommaso, per la sua costante disponibilità a confrontarsi e a stampare tutto quello di cui avessi bisogno.
Ad Andrea, amico e collega, per avermi supportato e sopportato nelle giornate più difficili. A Giacomo, per aver saputo gestire la mia ansia anche 24 ore su 24. A Giulia, per aver condiviso con me esperienze indimenticabili. A Marouan e Mattia, amici ingegneri, per avermi insegnato che Python e il Terminale non sono da temere.
Ad Alice, perché lei c'è sempre stata.
A Mattia, per la sua dolcezza e per avermi aiutato a dare spazio e forma alle mie idee. Ad Andrea, Federica e Leonardo, perché è importante lavorare di giorno, ma anche stare bene a casa la sera.
Al Gruppo *Amarcord*, per le accoglienze durante le mie apparizioni nell'amata Genova. Ed infine a tutti gli amici conosciuti in questi ultimi tre anni che mi hanno fatta sentire a casa dal primo giorno.

Bisogna chiamare le cose con il loro nome, la paura del nome non fa altro che aumentare la paura della cosa stessa.

—Hermione Granger, Harry Potter e La Camera dei Segreti, 1998

10. Glossario

A

Approccio non supervisionato: tecnica di apprendimento automatico che fornisce al sistema una serie di input che esso stesso classificherà per cercare di effettuare ragionamenti e previsioni.

Approccio supervisionato: tecnica di apprendimento automatico che mira ad istruire un sistema informatico.

B

Bag of words: insieme di parole che compone il dizionario di termini grezzi estratti dai documenti.

Barchart : tipo di grafico a barre la cui lunghezza è proporzionale al valore che rappresentano.

Beesawarm: modello che visualizza i dati come punti su un asse e che si espandono lungo l'altro asse per mostrare volume o quantità.

Breakpoints: punto di rottura, di demarcazione.

Bubble chart: tipo di grafico che mostra due o tre dimensioni di dati.

Bumpchart: modello visivo simile allo streamgraph utile per vedere variazioni nel tempo di diverse categorie. Rispetto allo streamgraph la bumpchart dispone secondo ranking ascendente o discendente sull'asse y ogni flusso rispetto ad ogni valore sull'asse x.

C

Clusterizzazione gerarchica: metodo di raggruppamento di elementi a coppie di due.

Computer science: serie di grafici simili che usano stessa scala e assi.

Corpus, corpora: un insieme di documenti.

D

Dashboard: letteralmente *cruscotto*, ovvero un sistema di visualizzazione frammentato in cui singoli elementi contemporaneamente visibili mostrano aspetti diversi dei dati a disposizione al fine di creare una visione olistica del dataset.

Data science: la scienza che studia le modalità di estrazione della conoscienza a partire dai dati.

Dendrogramma: modello utile per visualizzare relazioni gerarchiche o classificazione, Un esempio comune è l'albero genealogico di una famiglia.

Digamma: parola composta da più di un lemma.

Distant-close reading: lettura superficiale e lettura ravvicinata di un sistema complesso.

Dynamic Topic Models: algoritmi che consentono di analizzare l'evoluzione di topics estratti da un corpus di documenti distributi nel tempo

F

Force atlas layout: layout derivato dal modello più generale *force directed graph* presente nel software opensource *Gephi*.

Force directed graph: grafi in cui la distanza tra i nodi è il risultato di un parametro che li accomuna.

G

Glifo: rappresentazione astratta di un grafema, di più grafemi o di parte di un grafema, senza porre attenzione alle caratteristiche stilistiche.

H

Heatmap: modello visivo usato per mostrare variazioni di valore all'interno di una matrice a doppia entrata.

I

Inferenza: dal latino *inferre*, letteralmente significa portare dentro ed è il processo attraverso il quale da una proposizione assunta come vera si passa a una seconda proposizione la cui verità è derivata dal contenuto della prima.

Information design: ampio settore del design che si occupa della rappresentazione e della comunicazione visiva delle informazioni.

Information retrieval: settore dell'informatica che si occupa di creare sistemi efficaci per il recupero di informazioni specifiche all'interno di una grande quantità di dati

J

JSON: (JavaScript Object Notation) è un formato adatto all'interscambio di dati fra applicazioni client-server, basato sul linguaggio Javascript.

K

Keyword: parola chiave ed elemento più piccolo nella topic detection.

L

Latent Dirichlet Allocation (LDA): secondo questo metodo le parole di topic modeling i documento sono realizzazioni di misture di argomenti (topics). In particolare, ogni argomento si concretizza in una distribuzione di probabilità di tipo multinomiale su un vocabolario prefissato, l'argomento è compatibile con ogni documento del corpus ma la sua frequenza all'interno del documento varia statisticamente tra i documenti.

Latent Semantic Analysis (LSA): metodo per stimare il significato di un documento, la sua vicinanza a un particolare argomento o materia. In questo metodo di analisi sono prese in considerazione solo le parole che compaiono in almeno due documenti. L'analisi dell'LSA opera sotto la premessa che esiste una certa struttura semantica latente che è parzialmente nascosta dalla casualità della scelta delle parole.

Lemma: la forma di citazione di una parola in un dizionario.

Lemmization :processo algoritmico che determina automaticamente il lemma di una data parola.

M

Machine learning: campo della *computer science* che da ad un computer la capacità di imparare senza essere specificatamente programmato.

Matrice di similarità: una matrice sparsa che mette a confronto coppie di elementi.

Matrice sparsa: una matrice con diagonale nulla è il risultato dell'incrocio della stessa lista di valori. Quando i due valori omologhi si incrociano il risultato è zero.

Metaball: rappresentazione visiva del processo di distaccamento fluido di due elementi circolari.

Metadati: informazione, sotto forma di dato, che descrive un insieme di dati.

Modularity: calcolo algoritmico applicabile in Gephi che identifica clusters di nodi e permette di differenziarli cromaticamente attraverso l'interfaccia del software.

N

Named Entity Recognizer (NER):metodo che cerca di allocare e classificare i nomi di persona e città.

Nodo ponte: in inglese, "bridge node" è un nodo che connette due cluster.

Non negative matrix factorization: metodo per generare topic che si basa sulla fattorizzazione di matrici.

P

Pre processing pipeline: insieme di passaggi necessari a trasformare i dati grezzi in dati analizzabili da algoritmi.

Probabilistic topic modeling: metodo di topic modeling il cui obiettivo è quello di scoprire ed etichettare grandi quantità di documenti secondo il principio della probabilità.

Q

Query: in informatica, indica il termine che l'utente sta cercando all'interno di un database.

S

Sankey diagram: il diagramma di Sankey è un particolare tipo di diagramma di flusso in cui l'ampiezza delle frecce è disegnata in maniera proporzionale alla quantità di flusso.

Scatterplot: modello visivo che consente di visualizzare i dati in uno spazio x,y secondo la logica della dispersione.

Seed node: il nodo(i) della(e) pagina(e) principali che originano il grafo
Small Multiples:

Stemming: processo di riduzione della forma flessa di una parola alla sua forma radice.

Stop word: termine irrilevante all'analisi testuale: avverbi, articoli, pronomi, congiuzioni.

Streamgraph: tipo di grafico ad area disposto attorno ad un asse centrale. Appare come una forma organica e fluttuante.

T

Tensore: concetto astratto che costituisce una naturale estensione di quello di vettore.

Text mining methods: insieme di metodi per l'analisi testuale.

TF-IDF: funzione utilizzata in *information retrieval* per misurare l'importanza di un termine rispetto ad un documento o ad una collezione di documenti

Tokenization: analisi dei lessemi presenti in una frase e assegnazione di significato.

Tomografia Computerizzata (TAC): strumento di metodica diagnostica per immagini, che consente di riprodurre sezioni o strati corporei del paziente ed effettuare elaborazioni tridimensionali osservare il corpo umano al suo interno attraverso sezioni longitudinali o traversali.

Topic: gruppo di parole che tende a presentarsi sempre assieme all'interno di diversi documenti.

Topic detection: processo di rilevazione dei topic.

Topic modeling: tecnica statistica di analisi testuale che estrae gli argomenti astratti che emergono in una collezione di documenti.

Topic models: insieme di diversi algoritmi e calcoli statistici che effettuano il topic modeling

Treemap: un metodo per mostrare dati gerarchici usando rettangoli innestati.

11. Indice delle tavole