La speciazione del dato digitale

Alessandro Trivilini, docente e ricercatore al Dipartimento tecnologie innovative della SUPSI

Prendendo in prestito dalla botanica il significato del termine 'speciazione', è possibile intraprendere un breve viaggio alla scoperta dell'evoluzione del dato digitale negli ultimi vent'anni. Un ciclo di sviluppo importante caratterizzato da un processo evolutivo digitale che soltanto oggi, dopo molto tempo, desta l'attenzione e la perplessità di molti cittadini, che, sorpresi, iniziano a porsi domande su come sia stato possibile arrivare al punto che "Facebook sa cosa pensi, Google predice cosa cerchi, WhatsApp identifica con chi parli e Amazon condiziona il modo con cui spendi il tuo denaro".
Come siamo giunti a una 'profilazione' di massa dove un insieme di algoritmi riesce a persuadere le persone con una precisione emotiva millimetrica? Per trovare la risposta a questa e ad altre domande, che toccano da vicino la speciazione del dato digitale, è doveroso fare un passo indietro, per capire gli snodi principali attorno ai quali si è compiuta una serie di mutazioni che oggi portano a classificare il dato digitale come l'oro del nuovo millennio.

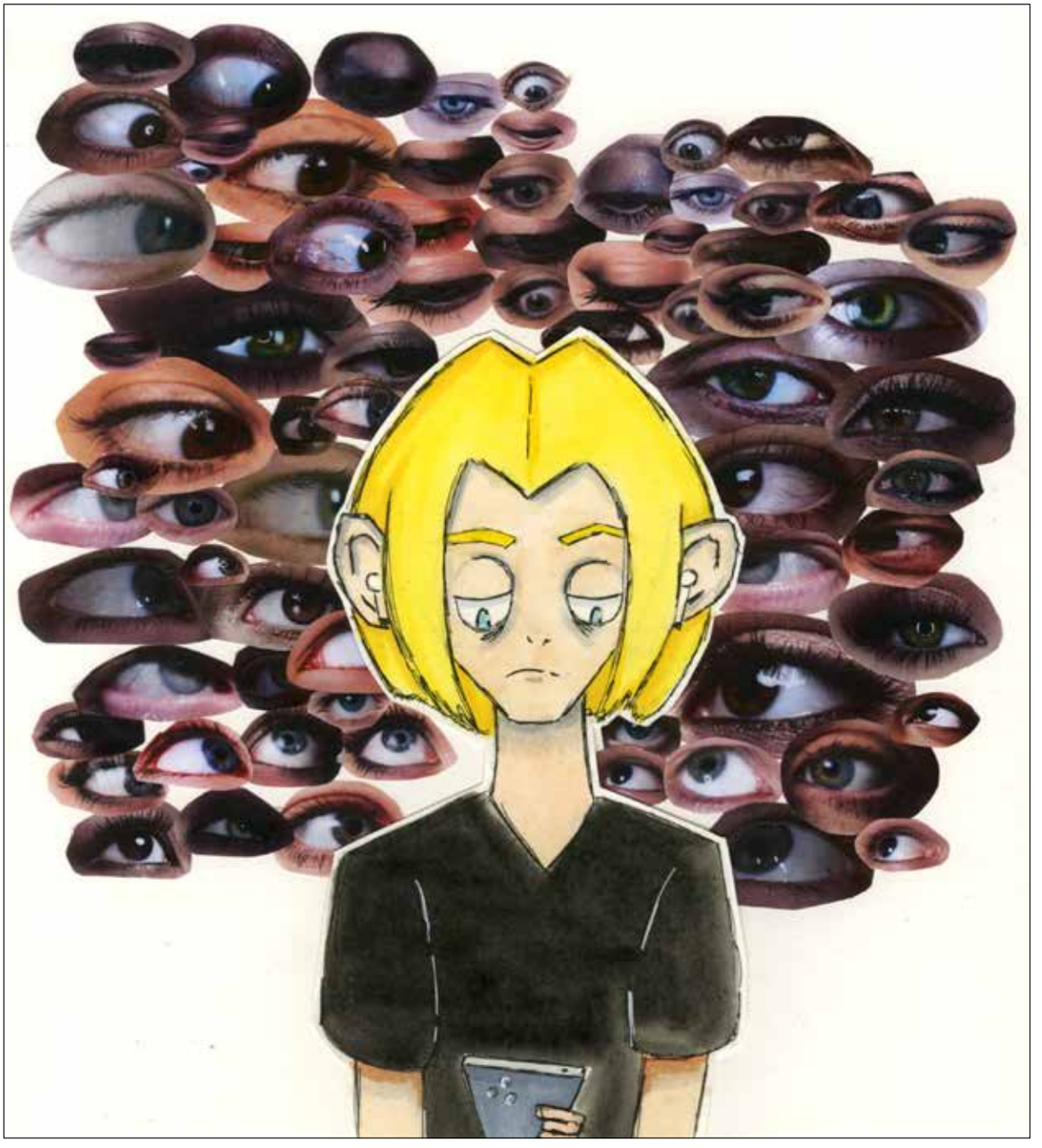
Lo snodo dei dati strutturati

Partiamo dalla fine degli anni Novanta del secolo scorso, quando l'uso della rete Internet da parte delle persone comuni stava per raggiungere il suo decimo anno di età. Un periodo storico molto importante, perché ha coinciso con un cambio di paradigma nei confronti della Rete, che ancora oggi caratterizza e condiziona il nostro comportamento quando cerchiamo le informazioni online. A quell'epoca le informazioni che si potevano trovare con i motori di ricerca di prima generazione, come Altavista, Hotbot, Excite, Arianna, Virgilio e Yahoo, erano memorizzate in un classico database, perché la loro quantità e la loro forma estetica permetteva di gestirle tranquillamente e ordinatamente nei singoli cassetti di una banca dati relazionale. Tutto era una novità, per cui tutto andava bene.
Infatti, il modello che ha contraddistinto questa fase tecnologica identifica il dato digitale come 'dato strutturato'. In pratica, ogni volta che un utente esprimeva una richiesta attraverso uno dei motori di ricerca appena citati, si innescava dietro le quinte, tra *client* e *server*, un comando specifico (*query*), che in modo preciso e mirato cercava tra i cassetti del database l'esistenza o meno del dato desiderato. Da un punto di vista tecnologico, in rapporto al numero limitato di dati diffusi in Rete, si può dire che la complessità tecnica di gestione era tutto sommato contenuta.

Ma la prima accelerazione non si è fatta attendere. Col passare del tempo un numero crescente di persone ha avuto accesso alla Rete, contribuendo così a far lievitare la quantità delle informazioni disponibili nel mondo virtuale. Si è trattato di un'evoluzione cruciale, che ha decretato ufficialmente la fine dei motori di ricerca di prima generazione, per dare spazio – ampio spazio – a qualcosa di completamente nuovo e travolgente.

Lo snodo delle informazioni non strutturate

Il 2001 è stato l'anno dell'arrivo di Google in Europa. Come un soccorritore fa con una persona in difficoltà, Google ha permesso al mondo intero, grazie al suo motore di ricerca, di continuare a cogliere le opportunità di Internet senza introdurre barriere tecniche dovute alla crescente dimensione dei contenuti. Il suo approccio e la sua efficacia ci hanno dato l'illusione che cercare gratuitamente informazioni in Rete fosse un diritto imprescindibile. Google ha gestito con cura il processo di speciazione del dato digitale, assicurandosi di dare il giusto benservito a chi, in precedenza, ha avuto l'arduo compito di aprire le porte alle ricerche online. La sua spinta innovativa ha classificato per sempre i dati strutturati, e ha dato vita alle informazioni non strutturate. Niente più database in cui memorizzare i dati diffusi in Rete dagli utenti, ma informazioni di ogni forma e genere contenute nelle pagine web, secondo il principio della popolarità.
Questo è stato il vero grande cambiamento di schema che ancora oggi ci supporta in ogni attività quotidiana, quando gestiamo ogni tipo di contenuto che possa essere digitalizzato, diffuso o condiviso in Rete. Con esso è cambiato anche l'approccio alla creazione e alla gestione dei contenuti multimediali. Il Web 1.0, in cui soltanto la figura del webmaster aveva le capacità tecniche per pubblicare in Rete, ha lasciato il posto al Web 2.0, ossia al paradigma in cui chiunque, senza competenze tecniche e manuali d'uso, può generare, gestire e diffondere in Rete contenuti multimediali con l'ausilio di nuovi strumenti tecnici, costruiti per essere semplici, multipiattaforma e pragmatici.
Con l'avvento di Google, oltre a un forte monopolio, è iniziata l'era dell'Internet senza limiti. Quanto più le parole contenute in una pagina web pubblica erano, e sono tuttora, ricercate, apprezzate e condivise da altre pagine web, tanto più Google le considera popolari.
Questo motore di ricerca è ancora oggi il più diffuso e utilizzato al mondo. Per questo motivo, vale la pena di



Giada Balinzo
2° anno di grafica – CSIA

conoscere, attraverso parole semplici, le quattro componenti fondamentali che caratterizzano il suo *modus operandi*.

La prima componente si occupa della gestione delle parole chiave che l'utente inserisce all'interno del riquadro adibito alle ricerche, sia dal punto di vista linguistico sia da quello matematico. Possiamo simpaticamente chiamarlo il 'gestore dei termini di ricerca'. Le regole usate dietro le quinte per interpretare la logica con cui l'utente inserisce le parole di ricerca sono definite da un modello e sono riconducibili all'algebra booleana. La sequenza con cui scriviamo le parole è importante, e in funzione dell'ordine di inserimento cambiano i risultati. Il loro concatenamento attraverso specifici operatori è la logica che usa questa componente per valutare in tempo reale la qualità delle ricerche e la loro concreta attuazione. Ovviamente, più parole inserisco, a prescindere dal loro significato, e più lo spettro di ricerca si riduce e, verosimilmente, meno risultati potrò ottenere. Per esempio, se inserisco le parole "Alessandro and Trivilini", otterrò risultati diversi da "Alessandro or Trivilini". Nel primo caso, il sistema imposterà le ricerche in tutti i documenti (pagine web) da esso indicizzati in cui le parole inserite compaiono forzatamente insieme. Nel secondo caso, invece, le ricerche vengono estese a tutti i documenti indicizzati (pagine web) in cui i termini inseriti compaiono singolarmente, portando così ad una unione dei risultati maggiore. Ogni ricerca effettuata dall'utente è codificata e analizzata dal sistema attraverso un albero binario che consente la sua rappresentazione formale e univoca.

La seconda componente è l'indicizzatore. Esso corrisponde a un archivio digitale potenzialmente infinito, nel quale vengono indicizzati i termini (parole) più popolari trovati in Internet rispetto alle pagine web che li contengono. Il principio dell'indicizzatore è che una pagina web di forte interesse non viene memorizzata nei grandi server di Google, ma indicizzata attraverso il suo indirizzo web e i termini (parole) che contiene. Di fatto, l'importanza di questa componente è strategica ai fini di un efficace ed efficiente reperimento delle informazioni, e il suo compito è oneroso e senza tregua. L'indicizzazione è un processo che una volta avviato non si ferma più. Lavora a stretto contatto con la prima componente (la gestione), dalla quale provengono le parole (termini) che l'utente inserisce nel motore di ricerca. Fra le sue peculiarità vale la pena di citare il processo di 'pesatura dei termini', ossia quella procedura altamente strategica che attraverso algoritmi molto complessi consente di assegnare un peso variabile ad ogni termine (parola) nel momento in cui viene identificato all'interno di una determinata pagina web. Segnalare semplicemente che un termine è presente in una pagina risulterebbe banale e poco interessante, mentre farlo attraverso un peso dinamico, che stabilisce il suo vero valore rispetto agli altri termini indicizzati, diventa molto interessante e redditizio.

Senza entrare in speculazioni, che di fatto non porterebbero a nulla, risulta comunque interessante osservare quanto diventi cruciale la generazione e la manipolazione dei pesi associati ai singoli termini. In questo modo, come per magia, una pagina web, per quanto rilevante e popolare possa essere, potrebbe non essere mai visualizzata sullo schermo dall'utente, se le venisse associato un peso molto basso. Al contrario, pagine web poco rilevanti potrebbero improvvisamente comparire fra i primi posti della classifica dei risultati. È una semplificazione, ma utile per conoscere cosa avviene dietro le quinte.

Una delle strategie usate per la generazione del peso di ogni termine prende in considerazione due fattori: la frequenza di ogni termine in ogni pagina web (quantità) in rapporto alla frequenza massima di quel termine in tutte le pagine web disponibili; la specificità di un termine e la sua precisione (qualità) per ogni singola pagina web. Questo approccio considera i termini meno presenti nelle pagine, consentendo così al sistema di ottenere un peso adeguato da associare ad ogni termine. Una sorta di bilanciamento per evitare che i furbi possano ingannare gli algoritmi di valutazione.

Passiamo alla terza componente, sicuramente quella più difficile da descrivere, poiché ricca di formule e di algoritmi matematici: il meccanismo di controllo. Possiamo definirlo il 'cervellone dell'intero sistema', colui che ha il compito di stilare la classifica finale dei risultati che il motore di ricerca in tempi brevissimi mostra sul video dell'utente. Le sue interazioni maggiori sono con la prima componente, dalla quale riceve i termini di ricerca che l'utente ha inserito in tempo reale, e con l'indicizzatore quale detentore assoluto del patrimonio di informazioni digitali precedentemente elaborate e indicizzate. Dispone di una quantità tale di algoritmi matematici capaci di apprendere autonomamente e di far rabbrividire anche il più talentuoso degli scienziati. A esso spetta il compito di codificare e modellare tutto ciò che l'utente chiede attraverso le parole chiave, e di stilare in tempi supersonici la famosa *ranking list*, ovvero la lista dei risultati con le pagine web più rilevanti per l'utente finale.

Infine, c'è una quarta e ultima componente principale: il *crawler*. Si tratta di un particolare programma informatico appositamente addestrato per la ricerca di informazioni rilevanti all'interno della rete Internet. Il suo compito consiste nel girovagare costantemente per la Rete in cerca di pagine web i cui contenuti sono molto visualizzati, cliccati o scaricati (cioè popolari). Un vero segugio digitale capace di stanare pagine web di ogni tipo. L'unico requisito è che esse siano popolari e fortemente referenziate da altre pagine web. Insomma, che abbiano un'ottima reputazione digitale. Quando un *crawler* identifica con precisione una pagina web che ritiene popolare, cerca di entrare in possesso dei suoi metadati, memorizzati all'interno dell'intestazione della pagina. Si tratta di informazioni descrittive come il titolo, gli autori, le etichette sul genere dei contenuti e altre informazioni utili. A tal proposito, vengono analizzati i contenuti della pagina in questione attraverso due fattori iniziali: la ricorrenza delle parole nella pagina e lo stile grafico con cui esse sono state espresse e pubblicate.

Queste sono le quattro componenti fondamentali che per vent'anni hanno silenziosamente lavorato per soddisfare tutte le nostre richieste, senza mai mostrare un segno di cedimento tecnico.

Lo snodo dei social newtork

Successivamente all'arrivo di Google hanno fatto la loro comparsa i social network. È senza dubbio una conseguenza naturale di un'evoluzione antropologica che ha trovato nelle informazioni non strutturate un forte alleato, oltre che un interessante stimolo.

I social network hanno dato alle persone una nuova dimensione virtuale in cui esprimere sé stesse. Tecnicamente facile da usare, e caratterizzata da una visibilità e popolarità pressoché illimitate, questa nuova dimensione, permettendo alle persone di condividere emozioni, gusti e comportamenti, si è affermata anche grazie alla forza propulsiva delle informazioni non strutturate.

Ancora una volta il processo di speciazione dei dati muta ed evolve, portando alla creazione di un nuovo continente digitale globale chiamato Facebook, popolato da oltre due miliardi di persone, che senza alcuna barriera culturale, tecnica, linguistica o ideologica diffondono spontaneamente le proprie informazioni sensibili.

Google e Facebook danno silenziosamente vita a un nuovo modello di business in cui il prodotto siamo noi, e la privacy è la conseguenza del nostro comportamento in termini di responsabilità e di consapevolezza nell'uso delle tecnologie digitali. Una situazione tanto semplice, quanto spietata.

Da poco si è concluso un ciclo di venti anni e questi colossi hanno raccolto dati che partono dall'ecografia per proseguire senza fine con i ricordi *postmortem* del proprietario dell'identità digitale social. Una macchina affascinante ma anche senza freni, che ha fatto della raccolta dati un modello tecnicamente inarrestabile, che a sua volta ha portato al più grande sistema di profilazione di massa che l'essere umano abbia mai conosciuto.

Un semplice blog come Facebook ha saputo spremere fino in fondo il valore delle informazioni, per capire quanto fosse volubile l'essere umano di fronte alla ghiotta opportunità di esprimere sé stesso oltre i fatidici quindici minuti di notorietà pronosticati da Andy Warhol. Il tocco creativo che ha decretato il successo di questa ulteriore accelerazione digitale risiede ancora una volta nella forza dei dati, dei nostri dati, unici e preziosi.

Per molto tempo la privacy sembra non essere mai esistita, nonostante ancora oggi in molti dimentichino che tutto ciò è stato possibile grazie al nostro consenso implicito, fornito al momento della creazione dell'identità digitale che ci permette di esistere all'interno di diversi social network. Ecco, quindi, che la pesca a strascico dei nostri dati personali, per come l'abbiamo conosciuta, potrebbe avere i giorni contati. È infatti in atto un cambiamento senza precedenti, e l'uomo deve farsi trovare pronto.

Lo snodo della privacy

Nonostante i problemi imposti dalla pandemia, da un punto di vista tecnico e tecnologico emergono all'orizzonte importanti novità che influenzeranno ulteriormente il processo di speciazione del dato digitale.

Tra queste spicca la nuova Legge federale sulla protezione dei dati (LPD), approvata dall'Assemblea federale nel settembre del 2020. Si tratta della revisione totale di una legge che guarda esclusivamente, e con particolare attenzione, ai dati personali delle persone fisiche. Come sempre accade in questi casi, quando entrano in gioco nuove regole, è auspicabile, oltre che dovuto, che si comprenda il dominio di applicazione delle stesse, onde evitare spiacevoli sorprese. In questo caso specifico, la prima cosa importante da conoscere e da non dimenticare è la definizione legale di 'dato personale', ossia di tutto ciò che in termini di informazione può identificare direttamente o indirettamente una persona fisica. Ad esempio, appartengono ai dati personali la data di nascita, il numero AVS, il numero di telefono, l'indirizzo e-mail, il numero IP fisso, il colore degli occhi, il peso e i tratti del carattere, oppure la religione, le opinioni politiche e tutte le informazioni sulla salute. Queste e altre informazioni sono soggette alla tutela della nostra privacy. Il diritto alla riservatezza è un diritto fondamentale e come tale è tutelato dalla nuova Legge svizzera sulla Protezione dei Dati, allineata a quella europea (GDPR).

Si apre quindi una nuova èra, in cui chiunque decida di trattare dati personali deve sottostare a regole chiare, pensate per offrire maggiore trasparenza, senza correre il rischio di limitare le innumerevoli opportunità offerte dalle tecnologie digitali del prossimo futuro.

Da tempo abbiamo chiesto nuove regole per la tutela della nostra privacy, ed ora sembra che il nostro desiderio sia stato esaudito. È tuttavia opportuno guardare oltre, e cercare di capire preventivamente quali potrebbero essere le possibili conseguenze di queste nuove regole, che, senza il buon senso e il principio di proporzionalità, rischierebbero di diventare troppo rigide e controproducenti.

Clara Prisco
2° anno di grafica – CSIA