L'influenza di Wikipedia

Gli epidemiologi vorrebbero prevedere le malattie come i meteorologi prevedono le piogge e sostengono che il modo in cui le persone navigano su Wikipedia potrebbe consentirlo.

The Physics arXiv Blog

Lo scorso anno, in questo periodo, il Centers for Disease Control and Prevention di Atlanta ha lanciato un concorso al fine di identificare il metodo migliore per prevedere le caratteristiche della stagione influenzale 2013-2014 utilizzando i dati raccolti da Internet. Kyle Hickmann e alcuni colleghi del Los Alamos National Laboratories in New Mexico, hanno rivelato i risultati del loro modello, che ha utilizzato in tempo reale i dati ricavati da Wikipedia per stimare le verità di base dei dati raccolti dal CDC, che emergono dopo circa due settimane.

I ricercatori sostengono che il loro modello ha la capacità di cambiare le previsioni sull'influenza, trasformando una specie di magia nera in una scienza moderna attendibile quanto le previsioni del tempo.

Ogni anno negli Stati Uniti muoiono tra le 3mila e le 49mila persone a causa di complicazioni legate all'influenza, per cui una previsione accurata può avere un impatto significativo sul modo in cui la società si prepara all'epidemia. Attualmente, il metodo utilizzato per monitorare lo scoppio dell'influenza è alquanto antiquato. Ci si affida a un sistema volontario in cui gli operatori della sanità pubblica riportano la percentuale di pazienti colpiti da sindromi simil-influenzali che vengono visitati ogni settimana; questa percentuale tiene conto delle persone con una temperatura superiore ai 38 gradi, tosse e nessun'altra ipotesi di malattia che non sia influenza. I numeri riflettono l'incidenza dell'influenza in ogni istante, ma la precisione è decisamente limitata. Non si tiene conto, per esempio, delle persone colpite dalla malattia che non seguono alcuna cura, o delle persone con sintomi simili a quelli dell'influenza, ma che in realtà non ne sono affetti.

C'è un altro problema importante. Il network che riporta i dati è relativamente lento e impiega circa due settimane per filtrare i numeri attraverso il sistema, per cui i dati risultano sempre vecchi di settimane. Per questo motivo il CDC è interessato a trovare nuovi metodi per monitorare la diffusione dell'influenza in tempo reale. Google, in particolare, ha utilizzato il numero di ricerche sull'influenza e i sintomi simil-influenzali per prevedere la malattia in varie parti del mondo. Questo approccio ha avuto un notevole successo, ma anche alcune carenze che lasciano perplessi. Un problema, per altro, è che Google non rende i suoi dati liberamente disponibili e in questo genere di ricerca una tale mancanza di trasparenza può creare problemi.

Così, Hickmann e il suo gruppo si sono rivolti a Wikipedia. La loro idea è che la variazione nel numero di persone che accedono agli articoli riguardanti l'influenza sia un indicatore della diffusione della malattia. E siccome Wikipedia rende i suoi dati liberamente disponibili a ogni interessato, la fonte risulta assolutamente trasparente e, probabilmente, rimarrà disponibile per il prossimo futuro. Hickman ha utilizzato i dati ricavati da articoli sull'influenza, risalenti agli anni precedenti, per preparare un algoritmo di apprendimento automatico che individui il collegamento con i dati di malattie simil-influenzali raccolti dal CDC. Successivamente, ha utilizzato l'algoritmo per prevedere in tempo reale i livelli della malattia durante la stagione influenzale.

I risultati costituiscono una buona base di dati, che il CDC rende disponibile dopo due settimane. «Gli accessi agli articoli di Wikipedia si sono rivelati altamente correlati agli archivi storici delle malattie simil-influenzali e permettono una previsione accurata dei dati alcune settimane prima che vengano resi disponibili», precisa Hickmann. Tuttavia, resta una precisazione da fare. Le previsioni sottovalutano in modo significativo la coda del periodo influenzale. Probabilmente, ciò succede perché le persone non sono solite tornare sugli articoli di Wikipedia che riguardano l'influenza se sono stati colpiti da un altro ceppo influenzale, che è una delle causa delle malattie stagionali.

Ciò nonostante, il lavoro costituisce un importante passo verso un sistema di previsione dettagliato e attendibile quanto le previsioni del tempo. Una caratteristica valida del metodo è che mostra quando il modello devia dai dati di base. Ciò permette di modificarlo in tempo reale e tenere conto di queste differenze, proprio come le previsioni del tempo. ■