Ricerche nel Web 2.0. Superati i sistemi semantici, adesso scende in campo la "patter recognition". Parola di Mike Lynch e Thomas Bayes, Abate del 18^ secolo

(Corriere Economia, dicembre 2007)

Quanti sono i documenti presenti nel web? Impossibile stabilirlo con esattezza, ma uno studio dell'Indiana University li stima attualmente in almeno 40 miliardi. Stiamo parlando di pagine online che all'interno contengono informazioni multimediali. Spesso non correlate tra loro. Perchè ai testi vengono associate fotografie e filmati. Un flusso informativo destinato a crescere in modo esponenziale con l'avvento del Web 2.0. Per intenderci siti alla YouTube e Flickr. Ma anche milioni di blog e forum. Insomma una Babele digitale in cui i tradizionali motori riportano risultati in eccesso. Così, alla fine consultiamo quelli delle prime pagine. «La ricerca online sta diventando molto più articolata della semplice digitazione di una parola chiave in uno spazio bianco. La complessità con cui abbiamo a che fare, riguarda i dati "destrutturati". Come ad esempio trovare frasi di senso compiuto all'interno di filmati o didascalie in una miriade di fotografie». A spiegare a Corriere Economia questa nuova problematica, che attende i cybernaviganti della rete, è Mike Lynch. Fondatore e Ceo di Autonomy, un'azienda inglese che progetta software e sistemi informatici per ricerche su Internet. Precisa: «non seguiamo però le metodologie di Google o Yahoo, con approfondimenti per gruppi di parole. Neppure usando procedure semantiche, cioè ricerche in sequenza dove si interpreta il significato delle parole».

Mike ha messo a punto un'idea originale, che viene da lontano. Lui ha conseguito il dottorato di ricerca al prestigioso Christ's College di Cambridge, con una tesi sulle reti informatiche e le teorie matematiche di Thomas Bayes. Il reverendo presbiteriano, primo a calcolare a metà settecento, le probabilità che un evento si verifichi partendo dalla frequenza con cui si è manifestato in precedenza. Un processo applicato oggi da Mike per mettere in relazione le informazioni "destrutturate" presenti nel web. Grazie a programmi basati su reti neurali che simulano il funzionamento del computer, riproducendo i modelli dei neuroni del cervello. L'obiettivo? Riconoscere le informazioni in base alle caratteristiche salienti. Autonomy, fondata nel 1996, lo chiama "pattern recognition" *(ndr. riconoscimento per forme).* Un esempio arriva dal sistema in uso dalla polizia inglese per svolgere indagini. Si parte dal reperimento delle impronte digitali, per trovare poi negli immensi archivi informatici i collegamenti con precedenti accuse e reperti. «Siamo riusciti a mettere in pratica le teorie matematiche del reverendo Bayes anche nel riconoscimento dei volti umani, partendo da alcuni tratti salienti - racconta Lynch - e applichiamo la stessa procedura software in campo letterario per cercare intere frasi all'interno di testi complessi presenti nel web».

Il software di Autonomy è in grado di riconoscere quadri e dipinti analizzando semplici particolari. Non solo. I ricercatori dell'azienda di Cambridge stanno lavorando a un progetto ambizioso: cercare informazioni all'interno di programmi televisivi e filmato (purchè in forma digitale). Senza essere costretti a etichettare ogni singola scena o frase, perché in questo caso occorrerebbero enormi potenze di calcolo e quantità di memoria inimmaginabili. Invece la strada percorsa da Mike è analoga a quanto fa un bambino che impara a parlare o apprende una lingua straniera, quando non sa ancora né leggere, né scrivere. «Perché il bambino contestualizza i vocaboli». Associandoli a momenti, situazioni ed espressioni dei genitori che li pronunciano la prima volta.